THE USE OF THE WORLD WIDE WEB IN EPIDEMIOLOGICAL RESEARCH

Alexandra Ekman

Stockholm 2006
Till mamma och pappa
ABSTRACT

The world is becoming smaller. Through technical innovations all kind of communication is simplified, people come closer, and borders are erased. The Internet plays a major part in this process not only through the simplification of everyday life - but also of scientific life. Increasingly laypersons access the immense information available on the Internet in search for health information. The scientific society, on the other hand, lags behind. It has yet to make use of the possibilities inherent in these technologies to their full extent. As populations become more dynamic and geographically dispersed, this becomes a necessity in order to collect information fundamental in epidemiological studies. As the outside world becomes smaller, the world of the Internet and its possibilities becomes larger.

This thesis and the studies therein fall within the field of e-epidemiology. This recently defined scientific field includes the acquisition, maintenance, and application of epidemiological knowledge and information using digital media such as the Internet, mobile phones, digital paper, and digital TV. In particular, we studied the dissemination of cancer risk sites and the collection of health information – both mediated via the Internet.

Despite the immense amount of health information available on the Internet, there are no guarantees as to the quality of the information accessible. In study I we performed a systematic search of the Internet and studied the quality of the found cancer risk sites. The results were discouraging, with few websites fulfilling the demands of the quality criteria recommended by the EU and with no improvements noted in the consecutive searches.

In study II-IV we used the Internet to collect health information from Swedish women in two large population-based studies of 50 000 and 25 000 women respectively, aged 18-60. We demonstrated the feasibility of using the Internet for data collection in large epidemiological studies. Level of education and income differed somewhat between the responders to the web and paper. Despite these differences, the bias from the association in the target population was similar for both response methods. As the absence of bias is important for the survival of the method, this was a very important find. Another advantage when using the web-mode is the possibility to study the lurkers (participants that enter, start responding to, but do not complete a web-questionnaire). As lurkers are potential responders, they present an important goal for research aiming at preventing drop out and increasing response rates. As such the results are another important argument for the endurance of the method. Although the world is getting smaller in the sense of simplified communication, the opposite holds true for the epidemiological society. The dynamic, mobile features of current populations complicate data collection for large population-based studies. The Internet holds advantageous properties with the capacity to overcome these problems.

The Internet access keeps increasing in all parts of the world and all parts of society. It is important that the scientific society realizes the need to distribute qualitative information so as to maintain the trust and faith of a growing number of Internet users. Concurrently the Internet is facilitating the globalization, thus complicating the collection of information from an increasingly mobile population. The scientific society needs to make use of this innovative tool with its possibility to erase borders and time zones, and capability to efficiently collect health information in epidemiological studies. The world is becoming smaller.
CONTENTS

1 LIST OF PUBLICATIONS 3

2 LIST OF ABBREVIATIONS 4

3 INTRODUCTION 5

3.1 NEW TIMES, NEW NEEDS 5
  3.1.1 E-epidemiology 5

3.2 INFORMATION DISTRIBUTION 6

3.3 INFORMATION COLLECTION 6

4 BACKGROUND 7

4.1 COMPUTER AND INTERNET -- ACCESS AND USAGE 7
  4.1.1 Sweden 7
  4.1.2 The World 8

4.2 THE DIGITAL DIVIDE 8

4.3 USE OF THE WORLD WIDE WEB 10
  4.3.1 Information dissemination 10
  4.3.2 Information collection 11
  4.3.3 Cost 12
  4.3.4 Time 12
  4.3.5 Quality 13
  4.3.6 Response rates 14
  4.3.7 Interactivity 15
  4.3.8 Sampling/samples/populations 15
  4.3.9 Security and ethical considerations 16

5 AIMS 17

6 MATERIAL AND METHODS 18

6.1 STUDY I - CANCER RISK SITES ON THE INTERNET 18

6.2 STUDY II AND III - WOMEN’S LIFESTYLE AND HEALTH COHORT 19

6.3 STUDY IV – WOMEN’S HEALTH COHORT 21

6.4 ETHICS 23

7 RESULTS 24

7.1 STUDY I - CANCER RISK SITES ON THE INTERNET 24

7.2 STUDY II – WOMEN’S LIFESTYLE AND HEALTH COHORT 25

7.3 STUDY III –WOMEN’S LIFESTYLE AND HEALTH COHORT 25

7.4 STUDY IV – WOMEN’S HEALTH COHORT 26

8 DISCUSSION 28

8.1 METHODOLOGICAL CONSIDERATIONS 28
  8.1.1 Selection bias 28
  8.1.2 Information bias 30
  8.1.3 Confounding bias 31
  8.1.4 Generalisability 31

8.2 FINDINGS AND IMPLICATIONS 32
8.2.1 Study I 32
8.2.2 Study II 33
8.2.3 Study III 33
8.2.4 Study IV 34
8.3 IMPLICATIONS FOR THE FUTURE 35
9 CONCLUSIONS 36
10 SVENSK SAMMANFATTNING 37
11 ACKNOWLEDGEMENTS 39
12 REFERENCES 41
1 LIST OF PUBLICATIONS

This thesis is based on the following papers;

I. Ekman A, Hall P, Litton J-E

II. Ekman A, Dickman PW, Klint A, Weiderpass E, Litton J-E

     Optimizing the design of web-based questionnaires – experience from a population-based study among 50,000 women. *Submitted*

IV. Ekman A, Sparén P, Dickman PW, Klint A, Litton J-E
    Feasibility of using the web for a population-based survey of correlates to HPV related disease in Sweden. *Manuscript*

*With kind permission of Springer Science and Business Media.*
## 2 LIST OF ABBREVIATIONS

<table>
<thead>
<tr>
<th>Abbreviation</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>ARR</td>
<td>Adjusted Relative Risk</td>
</tr>
<tr>
<td>AOIR</td>
<td>Association of Internet Researchers</td>
</tr>
<tr>
<td>BMI</td>
<td>Body Mass Index</td>
</tr>
<tr>
<td>CI</td>
<td>Confidence Interval</td>
</tr>
<tr>
<td>HPV</td>
<td>Human Papilloma Virus</td>
</tr>
<tr>
<td>ICT</td>
<td>Information and Communication Technology</td>
</tr>
<tr>
<td>IT</td>
<td>Information Technology</td>
</tr>
<tr>
<td>MI</td>
<td>Myocardial Infarction</td>
</tr>
<tr>
<td>RD</td>
<td>Risk Difference</td>
</tr>
<tr>
<td>RDD</td>
<td>Random Digit Dialling</td>
</tr>
<tr>
<td>RR</td>
<td>Relative Risk</td>
</tr>
<tr>
<td>STD</td>
<td>Sexually Transmitted Disease</td>
</tr>
<tr>
<td>URL</td>
<td>Uniform Resource Locator</td>
</tr>
<tr>
<td>www</td>
<td>World Wide Web</td>
</tr>
</tbody>
</table>
3 INTRODUCTION

About 500 years ago Gutenberg invented modern printing, thereby making text available to everyone. Not since then has a new technological innovation changed the possibilities for information exchange and communication so profoundly as the Internet.

3.1 NEW TIMES, NEW NEEDS

The interactive co-evolution of social, economic, cultural and technological trends, known as globalisation, affects the lives of people all around the world. Mediated through technological innovations, communication, whether it be travelling across the world for a face-to-face meeting, or using the Internet to send an e-mail, has reduced the distances in space and time between people, has made the world smaller. Regardless of where we are in the world, if the technological facilities are at hand, the local, national and international news are available, bills can be paid, schools applied to, courses taken, businesses performed, information sought, products bought, contact kept – all via the Internet. This trend can be observed in all age categories and by all parts of society; governmental as well as private sectors have realized the potential of this new mean of communication. We live in a highly interactive society.

As the traditional methods of communication such as paper and telephone are growing more and more inefficient, technologies such as the Internet, cell phones/text messaging, digital TV etc are increasingly used for communication across borders. The technologies are used for information retrieval and dissemination, at least in some parts of society. The modern person goes online to book a doctor’s appointment, then goes online to gather information regarding the medical condition and thus arrives at the physician with more knowledge regarding his/her condition and the potential treatment than the average patient a few years ago. Similarly, researchers within the medical sciences are learning to use the technical prerequisites present to rationalize the different parts of the research process. The efficiency of reference collection and information dissemination has dramatically increased and the international, electronic peer reviewing is enabled by means of the Internet. However, the scientific society has yet to make use of the possibilities inherent in these technologies to the full extent. For example, using the Internet for the collection of human data in epidemiological research has so far been sparse. In order to successfully collect lifestyle information pertaining to large, dynamic and geographically dispersed populations, faster and more efficient methods is a necessity.

3.1.1 E-epidemiology

In this thesis a few of the many different qualities offered by the Internet within the field of epidemiology were investigated (Study I-IV). These include the distribution and collection of longitudinal information regarding health and lifestyle factors. We define this area, known as e-epidemiology, as “the science underlying the acquisition, maintenance and application of epidemiological knowledge and information using digital media such as the Internet, mobile phones, digital paper, and digital TV. E-epidemiology also refers to the large-scale epidemiological studies that are increasingly conducted through distributed global collaborations enabled by the Internet”. Health care is an area that could benefit and change drastically with the implementation of such new technologies enabling more and easier communication for a growing part of the population.
3.2 INFORMATION DISTRIBUTION

Using the Internet as a source for health and lifestyle information is very common. As a consequence, the sources of information available increase. The quality of this information is not guaranteed however, and little is known about what is found by a layperson searching for health information on the Internet. In Study I, the information regarding cancer risk sites on the Internet was explored. Cancer was chosen as it is one of the most common causes of death in the industrialised world today (www.who.int - accessed 2006-09-22).

3.3 INFORMATION COLLECTION

The successful and systematic collection of data is central in any epidemiological study. The traditionally used methods such as face-to-face and telephone interviews as well as paper-questionnaires are increasingly failing to produce qualitative results within financially feasible limits. Tools that are better suited for the present dynamic populations are needed and the Internet presents a powerful alternative for the collection of data with several intrinsic features still unexplored. In Study II, III and IV we aimed at understanding the differences between this tool and the traditionally used paper-based method for collection of data in what were otherwise traditional epidemiological studies. This included differences between the persons using the different tools offered in a mixed-mode study, as well as aspects pertaining specifically to the web tool, such as questionnaire layout and design.
4 BACKGROUND

4.1 COMPUTER AND INTERNET -- ACCESS AND USAGE

A prerequisite for information collection and dissemination over the Internet is the access to a computer and the Internet, as well as the knowledge about how to search for information using these technologies. The number of computer and Internet users differs widely across the world, which is the most common criticisms of results generated using the Internet. However, the trend is positive all over the world with the number of users increasing steadily, which in turn is the single most important incitement for the continued and extended research using the Internet.

4.1.1 Sweden

The use of computers at a regular basis is very common in Sweden. The average number of users (at least once a week) is increasing among all age groups (16-74) and is more than 80%. It varies from nearly 100% among students to about 35% among female pensioners (Figure 1). More than half use the computer every day. Among the students about 90% use the computer at home, while among the pensioners the corresponding number is less than 50%. The second most common place where people use a computer is at work.

Figure 1. Proportion of persons aged 16–74 who used a computer at least once a week during the first quarter of 2005 by gender.
The use of the Internet follows the same pattern as that of computer use, approximately 83% use the Internet. Literally all adolescents but also a majority of the older are users. This is an increase compared to 2002-2004. The largest increases can be seen among the every-day users (45% in 2002, 58% in 2005) and among the oldest age groups. The difference between the genders is small, except in the oldest age groups where female users are less common.

4.1.2 The World

The countries with the highest number of Internet users in the world at present are New Zealand, Iceland and Sweden which all have an Internet penetration above 75%. Looking at the top 30 user countries in the world, the penetration is about 67% (http://www.internetworldstats.com/top25.htm, accessed 2006-09-20). On the other hand, there are still about 30 countries in the world with an Internet penetration below 1%. The trend of an increasing number of Internet users follows the same pattern all over the world with the largest increases seen in low-income regions, such as Africa, the Middle East, and Latin America/the Caribbean. In total the usage growth between 2000 and 2006 in the world was 200%, with about 16% of the total world population using the Internet (Table 2).

<table>
<thead>
<tr>
<th>World Regions</th>
<th>Population (%) of world</th>
<th>% Population (Penetration)</th>
<th>Usage % of World</th>
<th>Usage Growth (%) 2000-2006</th>
</tr>
</thead>
<tbody>
<tr>
<td>Africa</td>
<td>14.1</td>
<td>3.6</td>
<td>3.0</td>
<td>625.8</td>
</tr>
<tr>
<td>Asia</td>
<td>56.4</td>
<td>10.8</td>
<td>36.4</td>
<td>245.5</td>
</tr>
<tr>
<td>Europe</td>
<td>12.4</td>
<td>38.2</td>
<td>28.4</td>
<td>193.7</td>
</tr>
<tr>
<td>Middle East</td>
<td>2.9</td>
<td>10.0</td>
<td>1.8</td>
<td>479.3</td>
</tr>
<tr>
<td>North America</td>
<td>5.1</td>
<td>69.1</td>
<td>21.1</td>
<td>112.3</td>
</tr>
<tr>
<td>Latin America/Caribbean</td>
<td>8.5</td>
<td>15.1</td>
<td>7.7</td>
<td>361.4</td>
</tr>
<tr>
<td>Oceania/Australia</td>
<td>0.5</td>
<td>54.1</td>
<td>1.7</td>
<td>141.0</td>
</tr>
<tr>
<td>World Total</td>
<td>100</td>
<td>15.7</td>
<td>100.0</td>
<td>200.9</td>
</tr>
</tbody>
</table>

Table 2. World Internet usage and population statistics 2002-2006.

4.2 THE DIGITAL DIVIDE

The digital divide is a common concept used when discussing the Internet under many and different circumstances (Figure 2). It is the line separating the persons with access to and knowledge about information technology including computers and the Internet (the information “haves”), from those without (the information “have-nots”). As society changes due to the impact of the Internet, people without access are left behind, cut off from the many advantages offered by these technologies. These include both personal and professional development possibilities that an on-line population might benefit from. One can talk both about the international digital divide – the disparities between countries, as well as a domestic digital divide – the divide within countries. Although most if not all countries are increasing the use and access to Information and Communication Technology (ICT), the industrialised countries do so faster. Thus, the gap between the information “haves” countries and the information “have-nots” countries increases. The same holds true for the domestic digital divide – the information “haves” increase faster, thereby increasing
the gap. Socio-demographic variables such as age, income, education, and race have been shown to be the determining factors. As society becomes more dependent on the Internet, the digital divide, although not necessarily widening, could be said to be deepening. In countries such as the United States, federal, state, and local government together with private organisations are involved in ventures to narrow and eliminate the digital divide.

Figure 2. Internet users per 100 inhabitants by development status 1994-2004

The digital divide and the disparities that this implies are of importance in the context of epidemiological research, as it might lead to bias if the “have-nots” are not given a chance of participating in a study. As such, it is of great importance to take such potential differences into consideration when planning to implement epidemiological studies using the Internet as an information collection tool.

On the positive side, in countries with an existing IT-structure, the Internet may open the door for sub-populations that for different reasons might otherwise be isolated from society. In this sense the information technology could play a part in bridging the divides existing in our society. The elderly are the ones with the least Internet connectivity and access, and also the fastest growing population group in most developed countries. They potentially have much to gain by getting access to the Internet. This is a mean with which this group could enhance their independence as well as interact with society in an effective and easy way, also diminishing the exponential increase in health and medical services. Other groups benefitting from this new mean of communication are underrepresented populations such as people with stigmatized illnesses, for instance anxiety and depression. Such groups have been shown to prefer the Internet as a mean with which to access health information.
4.3 USE OF THE WORLD WIDE WEB

4.3.1 Information dissemination

The faltering trust accompanying information on the Internet could be the result of several facts; the lack of a natural quality screening structure, which to a larger extent exists for books, newspapers or TV, the immense information available, and the anonymity both for the information distributor and the searcher. Even though several information channels suffer from poor quality, the large quantity of available information and the easy access on the Internet increases the problem. With the implementation of quality seals and bullet points for self-checks for quality, several organizations are trying to control the quality of the information on the Internet. In spite of these efforts, most information remains unchecked.

Statistics from the USA give a picture of a nation that increasingly turns to the Internet for information, including health care information for which 79% of online Americans use the Internet (www.pewinternet.org - accessed 2006-09-12). This corresponds to about 93 million people, and makes looking for health or medical information one of the most popular activities online. In 2005, 12% of online adults claimed that the Internet played a crucial or important role as they helped another person cope with a major illness. Internet users search the Internet for health information at any given time of the day, research a diagnosis or a prescription, get tips from other caregivers and online patients about dealing with a particular symptom and give and receive emotional support. Despite the fact that their foremost trust among all information channels lies with the physicians, their first resort is the Internet. This information channel has empowered people, enabling them to ask more informed questions during appointments, which not always has affected the dynamic between doctor-patients in a positive way. The convenience of the medium, the breadth of the information available and the possibility of anonymity are highly valued features of the Internet as a source of health information.

The population that use the Internet as a source for health information includes a variety of persons such as patients, relatives of people that are ill, mothers, adolescents etc. All have different needs, needs that may well change over time, for example as the disease progresses, the child grows older, the start of the menstruation or menopause etc. There might also be differences in the information sought for, the way in which it is searched for and the trust that is bestowed on it. Established media and government Internet sites are considered most trustworthy with 79 percent of users saying that most or all information on those sites is reliable and accurate. However, it has been shown that people from different socio-demographic groups trusted different sources of online information. The fastest growing group of users of the Internet are Americans with the lowest incomes. Hence, it is important for the government, private organizations and the media to keep this in mind when distributing health information online and try to adapt to the different needs of all the different groups of information seekers. Their knowledge about how to search for information most probably differs, as the Internet access and skills vary. In a study using focus groups consisting of 11-19 year old adolescents in the UK and USA, the youths showed a variety of means by which to determine whether the information was trustworthy, such as looking at the Uniform Resource Locator (URL) (the USA adolescents recognised the .edu as a sign of a credible academic institution), avoiding...
personal home pages etc. The perception of a trustworthy source, however, varied. Similarly, a group of college-aged students (USA) were tested on their ability to find quality on a specific health topic online. Despite that their field of interest coincided with the topic of the search as well as the aim of the study, the search methods used were poor. Considering that a defined and structured search is more likely to give more specific information, these results are discouraging.

There is a certain amount of scepticism regarding the trustworthiness of the information available on the Internet. Several studies investigating the quality of health information online have been conducted. In a review of such studies, the study methods and rigor were found to be faltering and no conclusions were drawn. A longitudinal design was suggested as an alternative approach, which until then had only been applied in a limited number of studies with contradicting results.

4.3.2 Information collection

The search for a cost-efficient mean for data collection is constant and not unique to this time period. The data collection methods have changed during the last century – from face-to-face to telephone to paper methods – all to some extent due to cost- and efficiency reasons. Now, the time has come for the Internet to enter the scene of data collection tools.

How it started
The first documented mail survey took place in 1788 when Sir John Sinclair sent out questionnaires to all ministers of all parishes of the Church of Scotland. After 23 reminders he achieved a 100% response rate. Since then, the field of survey methodology has evolved greatly. In the last 25-30 years, the computer revolution has been a major driving force behind this fast progress, significantly improving the ability of sophisticated statistical methods. The medium was initially designed for scientific purposes, later used within the American army before it reached the homes of laypersons in the late eighties/beginning of the nineties. Within 4 years of its introduction 50 million people were users. In comparison it took 30 years to achieve this level of saturation for the radio and 13 years for the TV.

The research community has witnessed the same explosive development. Beginning in the late 1980s and early 1990s, the Internet, by way of e-mails and later the web, started appearing as a method of choice in surveys. Within the last couple of years, surveys previously using other traditional modes are now performed using the Internet.

Web-questionnaires
The Internet is a network of networks, a networking infrastructure connecting millions of computers together globally. The Internet and the World Wide Web (www or the web) is some times wrongly used interchangeably. The web is a way of accessing the information through the medium of the Internet. There are other ways, such as e-mail, instant messaging, Usenet news groups, etc., to disseminate information over the Internet (webopedia.com). In the studies described in this thesis, the web is the tool used although e-mail was used as a communication tool in one of the studies.
Factors often cited within the field of Internet surveys are cost, time, quality, response rates – all of which are important factors that have instigated studies aiming at comparing the possible advantage of Internet-methods (including e-mail, panels, web-questionnaires) to traditional methods.

4.3.3 Cost

Lowering of the survey costs is an often-cited advantage within the field of web-surveys. However, scepticism as to whether the method in fact does decrease the over all costs of a survey exists. One problem is the difference and often even lack of the reporting of the cost of each step of the study process, for example the fixed costs such as the construction of the questionnaire, and the variable costs, such as coding and analysing the results. Thus, an adequate comparison is difficult. As many of the steps involved in the design of the web-based study differ significantly from those involved in a paper-based study, this is necessary for a rational comparison of the two modes.

The initial set up is the major cost when choosing a web option. Therefore, the cost-effectiveness of a web-based survey to a large extent depends on the size of the study sample. Once a certain cut off point is reached, coinciding with the cost of the set up, the marginal cost associated with an increased sample size is low. Thus, today larger studies have more to gain, and as mentioned above, once the initial cost is paid, the marginal cost for additional participants is close to zero. The software needed for the design of a web-questionnaire has become easier to manage and has decreased in price during the last years making web-questionnaire design an affordable, more attractive alternative, even for smaller sized studies.

Several steps in the survey process are made unnecessary with a web approach, thus enabling cost savings. Printing, sending (postage to receiver as well as the return to the researcher), scanning, or data entry are unnecessary as well as personnel costs otherwise accompanying these steps in the study process. The data will, as a result of the intrinsic qualities of the web-questionnaire such as skip patterns, probability checks, and drop-down menus (see 4.3.5 Quality, for an explanation), need less data cleaning, again cutting personnel costs.

Depending on the design of the study the price can be cut even more, for example by using an e-mail invitation instead of sending an invitation by traditional mail. This is only possible however, when the e-mail addresses of the participants are known. It is hard to compare whether or not the web is a cost saving alternative since the cost for the different steps in the study design and logistics are seldom accounted for and the inconsistency between the studies are often large.

4.3.4 Time

The web mode is a time-efficient method for practically the same reasons as it is cost-efficient; fewer steps are needed in the survey process, and consequently time is saved. Nonetheless, this is partly a simplification, as in the case of costs it is dependent on which steps are included in the comparison. The construction of the questionnaire can be very
time consuming. Excluding this from the comparison makes it less accurate and less informative. On the other hand, if the study uses self-reporting on a web site as the mode of contact, time from the initial planning of the study to the reporting of the results, can indeed be saved \(^{40}\). Using a different design, however, is not clearly as time saving.

The time from data retrieval to analysis (i.e. data in a database) is shortened, as there is no need for scanning of questionnaires as when using paper mode. By using skip patterns, plausibility checks etc. (for an explanation of these techniques, see 4.3.5 Quality, below), the data cleaning step is made redundant, and again time is saved \(^{31}\). Today software exists, facilitating and accelerating the construction of web-questionnaires, a step that used to be time consuming. Whether or not the survey-fielding period (mode of contact and the time it takes, the mode of follow-up and the allowance for multiple contact) is shortened largely depends on the design of the study. Inexperience might reduce the time saves, either by unskillful use of the technical advantages (plausibility checks) or by, for example e-mail lists that are not up-dated, leading to a large amount of undelivered invitations and the necessity to use alternative means of invitation \(^{41}\). Increased delivery speed does not necessarily translate into a significantly shorter survey fielding period \(^{31}\).

### 4.3.5 Quality

By data quality we mean the quality improvements, which are a consequence of the technical prerequisites inherent in the tool. These include skip patterns that saves the participant from being presented with follow-up questions regarding issues of no interest to them, for example follow-up questions on smoking to a non-smoker. Another technical advantage is plausibility or range checks, which reacts on answers given outside acceptable limits, for example numbers outside the age span of the study population. Consistency-checks, verifying the logical relationship between two or more different answers (for example, a person stating that she is 12 years old and has given birth to 4 children is very unlikely) is yet another technical benefit. The two latter mechanisms make the respondent aware of the error, and give them the opportunity to correct it. The same is done when an item is missed. All these inbuilt checks aim at increasing data quality and facilitating for the respondent \(^{42}\). Initial results show that the item non-response is lowered and the amount of information given in open-end questions increases when using a web mode compared to paper mode \(^{43,45}\). Other studies show no differences between the web and paper \(^{46}\). Since the web-retrieved data requires less human transcription it is less error prone \(^{35}\).

Data quality also refers to the absence of bias caused by the desirability to be socially “correct” (taking social norms into account when responding), and that of giving answers to sensitive questions. There are preliminary results pointing at a decrease in socially desirable responses as well as an increased willingness to disclose sensitive information when using web-questionnaire compared to paper-questionnaires \(^{46,47}\). However, the results are inconclusive and the effect is thought to decrease as the use of web-based questionnaires increase \(^{46,48}\).

The mode effect on quality is important, in particular if the method is to be used in combination with other tools, in which case the different modes could have different effects. The results so far are inconclusive. Again, it appears to be a question of the design of the study and the population under study. For example, in a study of alcohol use among
a population of 3rd and 4th grade students, McCabe et al found no mode effect. Similarly, there were none found in a study among adolescents invited to answer a questionnaire regarding health in a web and a paper questionnaire, respectively. These populations are specific in that they are young and school bound, with an almost 100% Internet knowledge and access. Others have reported opposite findings.

4.3.6 Response rates

The response rate is commonly seen as an important and valuable proxy assessment of bias. With an increasing response rate, the risk or extent of non-response bias, or the possibility that the responders are not representative of the sample, decrease. Falling response rates is a problem because survey non-response is usually not random and in turn may bias survey results. Also, low response rates affect the precision of the results gained.

Fricker et al showed in a literature summary for Internet-based surveys that the response rates when using a web mode are fairly modest. In studies where the respondents have been given the choice of a web or a paper-questionnaire, a majority of the studies present results in favour of the paper option. The few studies with contradicting results were performed in samples with high Internet literacy. This could mean that as the Internet access and literacy increase in the population, the differences might decline or even cease to exist.

In a meta-analysis by Manfreda and Vehovar of web-surveys the overall completion rate varied from 1-95% with an average of 42% for 89 reported cases. The extreme variability is explained by the difference in sampling and design and a closer look at web-surveys with general invitation brings the completion rate to 25-28% (among 14 surveys). In another meta-analysis, Cook et al showed a similar average response rates, 39.6% (CI 1.2% -78.0%) among 68 surveys reported in 49 studies. Again, the variability of the response rates was large and again explained by the difference in sampling and design. Hence, a somewhat lower response rate is implied for web-based surveys. It is nevertheless important to remember the difficulty in comparing web-surveys since the literature thus far is rather scarce in comparison with that vast amount pertaining to paper, telephone and face-to-face surveys. It is not only a question of quantity but also quality – the web surveys that have been performed are more variable in their design and in the characteristics of the samples - making comparisons even harder. As such, the response rates gained will be very dependent on the context or population as well as on the design the researcher used for conducting web-surveys.

Paper-based surveys show a decline in response rates 1971-2000. Mean response rates among paper-based surveys published in medical journals is approximately 60%. Written and telephone reminders are associated with a 13% higher response rate compared with studies that do not use these methods, but there is a cost to be paid for reminders. For paper-based studies there is a lot of literature regarding ways to improve response rates. Pre-notification, individualised contact, multiple contacts, and incentives have all been shown to increase response rates in paper-based surveys. Some of these means of increasing response rates, such as a personalised salutation, pre-contacts and multiple
contacts are easily applied to web surveys. Preliminary results show that such means do increase web survey login and response rate \(^{28,54,58}\). Other means, such as incentives, prove harder to implement in web surveys, as they are hard to organize logistically, particularly if the web-survey is carried out using e-mail invitations \(^{59}\).

### 4.3.7 Interactivity

The Internet makes use of both audible and visual mechanisms to convey a message. Texts, sounds, videos or pictures can be used to facilitate the survey process, for example through memory aid. In addition to this, the Internet allows for interactive data capture with rapid checking of responses, at least at the form level \(^{37}\). Furthermore, the web allows for adjustments of unforeseen problems, removal or adding of questions or follow-up questions as new issues arise based on preliminary results. Instant feedback in the form of analysis of, for example nutritional intake after having filled-out a food questionnaire or summary statistics of individual responses given, are other advantages offered that could potentially increase response rates in web-based surveys \(^{60,61}\). Other interesting features of the web tool are real-time randomization of survey questions and/or answers and rotation of items, which offer new complex experimental designs for methodological research \(^{31,62}\).

### 4.3.8 Sampling/samples/populations

One of the major concerns when conducting a web-based survey is the risk of bias, as Internet access is not 100% in the study population, and there are differences between the “haves” and the “have-nots”. This will be dealt with in more detail in Methodological considerations, page 28.

The web has proven to be effective when the aim is to reach diverse, geographically dispersed as well as specifically targeted populations, which might be hard or impossible to reach with other modes \(^{63}\). Many new research opportunities present themselves as researchers now gain access to populations through web-portals, discussion groups and web communities. Computerized administration also allows researchers to obtain sample sizes that far exceed those obtained with most traditional techniques \(^{64}\). Mathy et al studied a population of lesbians and bisexuals, who are hard to reach using traditional methods. They constitute a rather small part of the population and would therefore demand extensive time and financial resources \(^{65}\). Other examples of the new uses of the Internet are the collection of sensitive data (sexual behaviour diary) from a population of university women, and structured treatment programs and preventive intervention for people with mental health concerns \(^{66,67}\). Another group for whom the web could potentially be an appropriate survey mode are deaf individuals \(^{68}\).

Non-probability sampling methods, such as volunteer panels of Internet users are increasingly being formed and used in the web survey industry \(^{69}\). By advertisement on well-visited sites, people are invited to register as panel members. Basic demographic information is collected at the first registration and the member is then given a password, thus preventing others than the registered member to answer. A well-known example is Harris Poll Online, with a panel of more than 6 million members from more than 145
countries (Harris Interactive website visited 2006-09-07). Weighting and scoring is used to draw samples claimed to be representative of the general population for collection of information on a variety of topics.

Another mean is to pre-recruit panels using probability-sampling methods such as random digit dialling (RDD) telephone surveys. The persons that agree to participate are sent an e-mail invitation including an URL as well as a personal identification and a passwords. Another method with the potential of obtaining probability samples from the full population, not just the Internet population is used by Knowledge Networks. They use RDD telephone surveys to invite people to participate. Following agreement, participants are provided with Web TV units and free Internet access. An advantage is that information about the non-respondents can be obtained at each stage, permitting detailed examination of likely non-response bias. The probabilities of selection are known at each stage, compared to the Harris Interactive approach allowing for weighting class adjustments as well as post-stratification.

4.3.9 Security and ethical considerations

To ensure future involvement, the trust of the participants is essential and therefore the protection of participant privacy, confidentiality and autonomy is vital. Due to the novelty of the web methodology, there is as yet no consensus for an ethical protocol. In 2002 the association of Internet researchers (AOIR), an academic association with an international scope promoting critical and scholarly Internet research across academic borders (www.aoir.org), issued recommendations regarding ethical decision-making and Internet research. Given the ethical pluralism, cross-cultural framework, and the technical prerequisites, specific considerations must be taken. This is particularly significant when using for example participants of a chat room. They should be informed of the study and the possibility of non-participation. In such a situation it is essential that the consent process (if one is necessary) is comprehensible for any possible visitor of that particular chat.

The difficulty in financing this type of innovative research also needs to be recognized as the major impediment it actually is. The perfect study, with a design solely for the purpose of testing the qualities and possible advantages of the web as a tool for data collection is yet to be done. As for now, we make do with the designs and study populations at hand, even thought the main purposes of these studies are not that of investigating the web as a tool for data collection. The studies II-IV included in this thesis should be considered in that context.
5 AIMS

- To study the cancer risk sites available on the Internet, and to estimate the quality of these sites and the possible change in quality over time - *Study I*.

- To study the feasibility of using web-based methods in population-based epidemiological studies – *Study II and IV*.

- To study the possible socio-demographic differences between web and paper responders and the effect that these differences may have on the choice of method – *Study II and IV*.

- To investigate how the order of the questions in a web-questionnaire affects the drop out and to characterise lurkers (participants that enter, start responding to, but do not complete a web-questionnaire) – *Study III*. 
6 MATERIAL AND METHODS

6.1 STUDY I - CANCER RISK SITES ON THE INTERNET

Survey procedures
In the autumn of 2001 we performed a search on the Internet with the aim of identifying cancer risk sites. These are sites that, after having answered a few questions regarding lifestyle and health, present the user with a risk of developing a particular cancer. In order to minimize the size of the search, only sites evaluating the risk of breast, prostate, colon, and lung cancer were included. Six well-established search engines and one Meta crawler (a Meta crawler allows you to search multiple search engines at once) were used for the search. In the autumn of 2002, exactly the same search procedure was performed to study a possible change in test site prevalence. In January 2005 a smaller search was performed to investigate if conclusions reached in 2001 and 2002 were still applicable in 2005.

A simple search using a single common word or phrase will result in thousands and even millions of hits. The use of a specific word is therefore an important means of making a search more effective. The following combinations of words were used in the search strategy:

“cancer risk” +questionnaire –study –“alternative medicine”
+breast/prostate/colon/lung

For a discussion regarding the reasoning behind the above chosen words, see Study I.

Identification and quality assurance
The first 50 consecutive and unique URLs on each search engine and the Meta crawler were studied. Only sites that provided direct links to the risk assessment tool were included. Tests for heredity/genetic susceptibility were excluded, as were inaccessible sites, link pages, non-English sites and sites with at most a peripheral relation to the cancer test in question. This, combined with the abundance of duplicate sites, accounts for the low number of test sites actually analyzed.

Evaluation variables
Websites were analyzed according to a set of drafted guidelines for the quality criteria for health related websites. These criteria were:

(i) **Transparency** - The address of the organization or person responsible for the webpage as well as the purpose and objective of the site/page must be stated. A clearly defined target audience and source of funding must be declared.

(ii) **Authority** - A clear statement of sources of information should be provided.

(iii) **Privacy** - There should be a clearly defined security and confidentiality policy and system

(iv) **Currency** - Date for and frequency of updating must be noted.

These criteria were drafted by the EU and were also used by the Swedish National Board of Health and Welfare when investigating the quality of Swedish health websites 2001. These quality criteria are included in the quality criteria used in several other available guidelines. The quality criteria we used were “process” oriented, in order to be useful for a lay audience.
The Women’s Lifestyle and Health cohort started in 1991 when a random sample of 96,000 women born between 1943 and 1962 residing in Uppsala Health Care Region (comprising about one sixth of the Swedish population) was selected from the Swedish Central Population Registry at Statistics Sweden. These women received an invitation to participate in a study by responding to a paper-questionnaire with just over 100 questions regarding lifestyle and health. 49,259 of the women responded (51.3%). This study was performed in collaboration with researchers in Norway.

In 2003, the women that responded to the 1991 questionnaire and were still alive were contacted again. This time they received an invitation via traditional mail including the URL to a study website and a personal login. They were encouraged to enter the study webpage, log in and answer a web-based questionnaire. The web-questionnaire included approximately 90 questions updating the information gathered in 1991. It took approximately 1.5 hours to complete.

Figure 3. Study flowchart – Women’s Lifestyle and Health, Study II.

A reminder was sent to the non-responding women, approximately 3 months after the original invitation was sent (Figure 3). The majority of the non-responders received a paper invitation with the URL to the study website whilst a random sample of 5000 of the non-responders received a paper invitation and a paper-questionnaire, thereby offering them the choice of paper or web-questionnaire. Roughly 1000 women had entered the study website, recorded their e-mail address but had not responded to the web-questionnaire. This group of women were divided into three groups; one was reminded via paper and given the choice of a paper or a web-questionnaire, the other two were reminded via e-mail and
offered to respond to a web-questionnaire, either through a direct or an indirect link in the 
email (Figure 3).

A final reminder was sent to the non-responders in November 2003. All non-responders 
who had not left an e-mail address were invited via traditional mail and offered a paper 
questionnaire. Approximately 1000 women, for whom we had e-mail addresses were again 
contacted via e-mail, which included an indirect link to the website. The study was 
terminated in February 2004 when the web-based questionnaire was taken off-line.

Two versions of the web-questionnaire were created, “Hard-to-easy” and “Easy-to-hard”. 
“Easy” questions were those with dichotomous or 3-5 response alternatives such as “What 
are your working hours?”. Questions perceived as difficult demanded recollection of for 
example a certain year, name or a combination of these. In the “Easy-to-hard” version the 
easier questions were put in the beginning of the questionnaire with the level of difficulty 
increasing toward the end of the questionnaire. In the “Hard-to-easy” version the order 
was reversed.

The responses to the questions in the web-questionnaire given by the women were not 
registered until they had reached the end of the web-questionnaire and had pressed the 
“submit” button. Irrespective of whether or not the women completed the questionnaire, 
the time spent on each web page was recorded. This enabled a comparison of the average 
time spent on a specific web page (which corresponded to a question/several questions) 
between the two versions of the questionnaire as well as the difference in drop out rate 
between the two versions.

The definitions used in study III (described below) are defined by Bosnjak et al. but 
modified to fit the prevailing circumstances. The type of operative system used by the 
participant was also recorded.

**Responder web**
Women who logged in on the study website and started answering 
the web-questionnaire (Figure 4). Some completed the web-
questionnaire, some dropped-out but subsequently answered a 
paper-questionnaire, some dropped-out and did not subsequently 
answer any questionnaire.

**Responder paper**
Women that never logged in to the study website but that responded 
to the paper questionnaire. (Not included in the analysis in study III)

**Non-responder**
Women that never logged in on the study website nor answered the 
paper questionnaire.

**Completer**
Women in the “responder web” group that reached the end of the 
web questionnaire and pressed the “submit” button.

**Non-completer**
Women in the “responder web” group that logged in, started 
answering the questionnaire but never pressed the “submit” button.

**Lurker**
Women in the “responder web” and “non-completer” group who 
logged in but never completed the questionnaire. These fall in to one 
of the sub-categories below:
- subsequently responders, paper
  Lurkers who later responded to the paper questionnaire.
- subsequently non-responders
  Lurkers who never answered any questionnaire at all, neither paper 
or web.
Figure 4. Women's Lifestyle and Health. Response rates by group. Study III.

6.3 STUDY IV – WOMEN'S HEALTH COHORT

Study population
In Denmark, Iceland, Norway and Sweden, a phase III vaccination trial against Human Papilloma Virus (HPV) is being performed. For this purpose 25 000 randomly selected Swedish women, aged 18-45 years, were invited to participate in a survey. Participants were sent an invitation by post inviting them to log on to the study website and answer a web-based questionnaire. The invitation was sent randomly in four batches with a couple of weeks in-between each. Women who did not respond were sent a reminder letter including a paper questionnaire. After another 3-4 weeks they were sent a reminder letter including a web-based questionnaire. Women who still had not answered after another 3-5 weeks were called on telephone and asked to participate in a telephone interview or answer by paper or web-questionnaire (Figure 5). The women in the fourth batch received a paper-questionnaire with the first invitation thus offering them the choice of a web or a paper questionnaire already in the original invitation. The questionnaire included 27 questions regarding life and health issues as well as Internet access.
Figure 5. Study flowchart including response rates, Study IV. The different yellow/orange colours denote the different send outs. Light grey = batch 1-3 = group A. Dark grey = batch 4 = group B. The number in *italic style* above the arrows are the response rates out of total within each batch. The numbers below the arrow are the response rates for the different response types within each response period.

We used a randomized design to divide the women into the batches (1-4) forming groups A and B. Thereafter self-selection in response patterns took place. This allowed us to study how socio-demographic variables affect the process of selection between modes (Table 4).

**Modes of response offered**

<table>
<thead>
<tr>
<th>Original send out</th>
<th>Reminder 1</th>
<th>Reminder 2</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mail invitation</td>
<td>Mail reminder</td>
<td>Telephone reminder</td>
</tr>
<tr>
<td>Batch 1</td>
<td>Batch 2</td>
<td>Batch 3</td>
</tr>
<tr>
<td>Web</td>
<td>Web</td>
<td>Web</td>
</tr>
<tr>
<td>GROUP A</td>
<td>Batch 2</td>
<td>Batch 3</td>
</tr>
<tr>
<td>Batch 3</td>
<td>Web</td>
<td>Web</td>
</tr>
<tr>
<td>Batch 4</td>
<td>Web/Paper</td>
<td>Web/Paper</td>
</tr>
</tbody>
</table>

Table 4. Definition of groups in Study IV

**Statistical analysis Study II-IV**

We calculated response rates (proportion responding), drop out rates (proportion non-completers) and estimated ratios of proportions (relative risks) to compare the response rates between subgroups defined by mode and type of responder. Comparisons of socio-demographic variable distribution across subgroups were performed. Due to the large sample size even small differences without practical relevance may be statistically significant.
In Study II and III the women had previously answered a questionnaire (1991/1992). This enabled us to estimate measures of association in the study population (women alive at the initiation of the follow-up 2002/2003). We compared these to estimates based on responses to the follow-up questionnaire. To estimate potential bias caused by non-response and method used for data collection (web or paper questionnaire), the relative risk for a number of known exposures (smoking and physical activity) and outcomes (myocardial infarction, parity and Body Mass Index (BMI)) were analysed.

In Study III non-parametric equality-of-median tests were performed to compare the medians of the time per web page between the different versions of the web-questionnaire. Differences in drop out rates were visualized using Kaplan-Meier curves. These were compared using the log-rank test for equality of survivor functions.

In Study IV we tested whether the response methods offered affected the total response rate. Selective response patterns were evaluated by comparing the distribution of socio-demographic variables between the web-responders and paper-responders, as well as the responders to the original send-out and the first reminder. We analysed the item non-response among two sets of questions; nine questions of a more general kind (background information, socio-demographic variables) and eight questions regarding sensitive issues such as number of sex partners and having been diagnosed with a sexually transmitted disease (STD) (see Appendix A in Study IV).

6.4 ETHICS

The Ethical Committee at Karolinska Institutet approved the study protocol for studies II-IV. The information contributed by the participants was handled confidentially and protected by the Personal Data Act 77. The participants were not asked to fill out their name, address, or personal identification number as these were known to us. The unique username served as an identifier. For the paper questionnaires in Study II-IV, a serial number was created that served as a unique identifier. For the paper questionnaires in Study II-III this number was printed as a number on each questionnaire, in study IV it was printed as a European Article Number-code.
7 RESULTS

7.1 STUDY I - CANCER RISK SITES ON THE INTERNET

The different search engines used for the search of cancer risk sites gave approximately the same result, i.e. the same websites were found with each search engine. This gave us reason to believe that our search strategy found most of the currently available cancer risk sites. The number of hits had increased by approximately 50% in the search performed in 2002 compared to 2001. This is in line with current statistics, which show a steady increase in health-related websites. In all, 12 cancer risk sites were found in the search performed in 2001, out of which 8 could still be found in the search in 2002. In addition, 10 new test sites were identified. In 2005 a search using only Google was performed identifying three websites with tests (one new, and two the same as found in previous searches). There was no improvement seen regarding the fulfilling of the quality criteria. The same search was performed also for breast cancer with the same results.

Transparency
Stating the purpose and objective was the criterion met by most of the websites. In total, three of the websites (out of 12 in 2001 and 18 in 2002) did not define either the purpose or objective. The target audience was stated in approximately half of the cancer risk sites both in the 2001 and 2002 search. In the search in 2001, an address (physical or electronic) for the organization or person responsible for the site was found for approximately a third of the sites. This had improved somewhat in the search in 2002.

The increase (in absolute numbers) in sites with addresses was thus due to new test sites found in the 2002 search with information about the address. None of the sites lacking an address in the 2001 search had added this information in 2002. The name or address of the person or organization responsible for the questionnaire itself was frequently hard to identify, and often required considerable searching. It was often easier to find the address of a responsible person or organization for cancer tests on larger Internet portals. Sources of funding (grants, sponsors, advertisers, non-profit, voluntary assistance) were found on two cancer risk sites in each of the searches (2001 and 2002). These websites were the same in both searches.

Authority
The sources of information were stated clearly for half of the cancer risk sites in the 2001 search and approximately half of the sites in the 2002 search. Again the increase (in absolute numbers) was due to new cancer risk sites incorporating this information, and not a consequence of an update of sites found in the 2001 search.

Currency
Information about updates and their regularity was very hard to find on the majority of the cancer risk sites. Less than a third of the test sites indicated when they were last updated. The regularity of these updates was not stated on any test site.

Privacy
None of the cancer risk sites required personal identifying data, which made the necessity of a privacy policy redundant.
7.2 STUDY II – WOMEN’S LIFESTYLE AND HEALTH COHORT

The overall and group specific response rates can be seen in Figure 3 on page 19. Responders were, in general, more likely to have a higher level of education and a higher income compared to non-responders, to smoke less and to be born in a Nordic country. No differences were seen between responders and non-responders with regard to age, BMI, physical activity level and parity. In general, the women responding to the web-questionnaire had a higher family disposable income compared to both the paper responders and the non-responders. No association was found between smoking, BMI, parity and level of activity and the probability of responding to a questionnaire, irrespective of whether it was a web or a paper questionnaire.

A comparison was made between the responders (web mode) and non-responders in group 1.2 (Figure 3). Women with >12 years of formal education were twice as likely to respond to the web questionnaire at this stage compared to women with <9 years education. Smoking and having lived in a Non-Nordic country for the largest part of your life were negative predictors of answering the web-based questionnaire. When comparing web responders from the first reminder (group 1.2, Figure 3) with the non-responders of the group that subsequently received and answered a paper questionnaire in the second reminder (pertaining to group 2.1), the strongest association was seen for the women with more than 12 years education.

The relative risk (RR) for myocardial infarction comparing smokers to non-smokers (adjusted for level of education) was compared between the source population (i.e. using data collected in the 1991/1992 questionnaire) the non-responders, the web-responders and the paper-responders. The estimated RR among the web-responders differed from that among the source population, but the difference was comparable to that estimated among the paper-responders (see Study II, Table 5).

7.3 STUDY III – WOMEN’S LIFESTYLE AND HEALTH COHORT

There were significant differences between the socio-demographic variable averages and proportions tested respectively. The lurkers (subsequently responders paper) differed significantly from the responders paper with respect to average age, proportion >12 years of education, number of children and BMI. Similarly, the lurkers (subsequently non-responders) differed from the non-responders with respect to education, BMI, activity level and proportion of the responders having lived a majority of her life in a non-Nordic country. There are no differences in IT prerequisites, measured as operative system used by the participant, between the completers, lurkers (subsequently responders paper) and lurkers (subsequently non-responders).

The risk difference (RD) for completion among completers of the “Easy-to-hard” version compared with completers of the “Hard-to-easy” version of the questionnaire was 0.06 (95% CI: 0.05-0.07).

Generally, the median time per page was shorter for a question when placed late compared to early in the questionnaire. The median time was also longer for pages with several questions as well as for pages with complicated questions, demanding recollection of times, dates and durations (see Study III, Appendix A).
Figure 6 shows that the drop out rate in the “Easy-to-hard” version was lower and significantly different from the drop out rate in the “Hard-to-easy” version of the questionnaire.

Figure 6. Women’s Lifestyle and Health, study III. Curve: Kaplan-Meier drop out estimates for the two versions of the web-questionnaire (“Hard-to-Easy”=dotted line, “Easy-to-Hard”= solid line). Graph: Number of dropouts (N= 2 212) at each web page for the two versions of the web-questionnaire.

7.4 STUDY IV – WOMEN’S HEALTH COHORT

The final response rates in groups A and B were equal (63% and 62%, respectively) (Figure 5). The distribution of socio-demographic variables among the responders in group A and B after the original send out differed significantly on all investigated variables except smoking. After reminder 1, there were no differences between groups A and B responders except for Internet use and civil status. Comparing the total number of responders after two reminders, there were no differences except for civil status, Internet usage and mode used for response. Although significant, these differences were minor.

Factors positively affecting the risk of being a web-responder was ≥13 years of education compared to ≤12 years, being a current smoker compared to never and women that were currently students or pensioners/un-employed or disability pension compared to full-time employees (Table 5).

The RR of responding to the web compared to paper questionnaire in reminder 1 group B decreased with age (P<0.001). The adjusted relative risk (ARR) was lower for part-time workers and pensioners/un-employed/disability pensioners compared to full-time
employees. Furthermore, ≥13 years of education compared to ≤12 years, and being a student compared to full-time employment increased the risk of responding to the web (Table 5).

Table 5. Adjusted relative risks of responding to the web compared to paper in group B, original send out and group A, reminder 1, Study IV.

<table>
<thead>
<tr>
<th>Variable</th>
<th>Web vs. paper, group B original send out</th>
<th>Web vs paper, group A, reminder 1</th>
</tr>
</thead>
<tbody>
<tr>
<td>Age</td>
<td>Adjusted relative risk*</td>
<td>95% CI</td>
</tr>
<tr>
<td>18-24</td>
<td>77 306 1.00 Ref</td>
<td>194 709 1.00 Ref</td>
</tr>
<tr>
<td>25-29</td>
<td>43 265 0.80 0.55-1.16 145 639 0.94 0.76-1.16</td>
<td></td>
</tr>
<tr>
<td>30-34</td>
<td>54 313 0.93 0.64-1.33 130 799 0.74 0.59-0.93</td>
<td></td>
</tr>
<tr>
<td>35-39</td>
<td>48 309 0.88 0.60-1.30 123 826 0.71 0.56-0.91</td>
<td></td>
</tr>
<tr>
<td>40-46</td>
<td>52 324 0.91 0.62-1.33 114 884 0.65 0.50-0.84</td>
<td></td>
</tr>
<tr>
<td>Civil status</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Married/co-habiting</td>
<td>196 1121 1.00 Ref</td>
<td>478 2780 1.00 Ref</td>
</tr>
<tr>
<td>Single</td>
<td>69 301 1.00 0.76-1.32</td>
<td>198 813 1.03 0.87-1.23</td>
</tr>
<tr>
<td>Divorced/Widow</td>
<td>9 95 0.55 0.29-1.04</td>
<td>30 264 0.77 0.54-1.09</td>
</tr>
<tr>
<td>Education</td>
<td></td>
<td></td>
</tr>
<tr>
<td>≤12</td>
<td>87 602 1.00 Ref</td>
<td>257 1755 1.00 Ref</td>
</tr>
<tr>
<td>≥13</td>
<td>187 915 1.38 1.08-1.76</td>
<td>449 2102 1.28 1.10-1.48</td>
</tr>
<tr>
<td>Smoker</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Never</td>
<td>144 862 1.00 Ref</td>
<td>429 2185 1.00 Ref</td>
</tr>
<tr>
<td>Past</td>
<td>69 69 1.07 0.81-1.42</td>
<td>132 759 1.01 0.84-1.21</td>
</tr>
<tr>
<td>Current</td>
<td>69 69 1.37 1.06-1.78</td>
<td>145 913 0.89 0.74-1.06</td>
</tr>
<tr>
<td>Employment</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Full-time</td>
<td>95 601 1.00 Ref</td>
<td>276 1543 1.00 Ref</td>
</tr>
<tr>
<td>Part-time</td>
<td>37 321 0.78 0.54-1.12</td>
<td>100 786 0.76 0.56-0.95</td>
</tr>
<tr>
<td>Self-employed</td>
<td>9 48 1.19 0.63-2.22</td>
<td>23 119 1.15 0.78-1.70</td>
</tr>
<tr>
<td>Home-worker/maternity</td>
<td>16 128 0.81 0.49-1.35</td>
<td>55 352 0.83 0.63-1.10</td>
</tr>
<tr>
<td>Student</td>
<td>76 253 1.51 1.10-2.08</td>
<td>187 552 1.34 1.11-1.63</td>
</tr>
<tr>
<td>Pensioner/un-employed</td>
<td>41 166 1.46 1.05-2.08</td>
<td>65 505 0.76 0.59-0.98</td>
</tr>
</tbody>
</table>

* RR of responding to web compared to paper (controlling for age, civil status, education, smoking, and employment.)

The ARR of responding to reminder 1 paper vs. original send out web group A was higher for the three oldest age groups compared to the youngest, for divorced/widows compared to married/co-habiting for part-time workers and home-workers/maternity leave compared to full-time employees (see Study IV, Table 5). It was lower among women with a higher level of education and students compared to full-time employers.

The ARR of responding to reminder 2 paper vs. telephone (groups A and B) was lower for singles compared to married/co-habiting and Internet users. It was generally higher for responders in group A compared to group B adjusted for civil status and employment.

The RR of any item non-response was investigated among responders to the original send out group B as well as reminder 1 responders group A. The RR of item non-response was higher among paper responders compared to web responders, irrespective of batch and response period although the RR found were generally larger in the original send out than in the reminder. The relative risks were not as large comparing paper and web responders for sensitive questions. Adjusting for socio-demographic variables generally increased the relative risk.
8 DISCUSSION

8.1 METHODOLOGICAL CONSIDERATIONS

The two major sources of errors occurring in epidemiological studies are random errors and systematic errors. Random error is the variability in the data that we cannot readily explain. This leads to a reduction of the precision of an estimate, indicated by a wider confidence interval. Larger study populations minimize random error. Systematic error includes selection bias, information bias and confounding. In survey methodology the terms used are coverage, sampling, non-response and measurement error. These must all be evaluated relative to the costs of conducting a high quality survey. A brief description of each of the mentioned errors and the bias that they might cause is presented in view of a web-based approach. The implications for studies I-IV will also be discussed.

![Diagram of study concepts and bias]

Figure 7. Overview of study concepts and the bias associated with these.

8.1.1 Selection bias

Selection bias stems from the procedures used to select subjects and from factors that influence study participation.

Coverage error - This occurs when the frame does not cover all units in a target population and there is a difference between uncovered and covered. For example, if the target
population is the Swedish population, the frame might be a telephone list of the Swedish population at a given time (Figure 7). As a 100% of the population does not have a traditional phone, the frame will not cover the entire target population and consequently everybody does not have an equal or non-zero chance of participating. If the covered and non-covered differ, a bias in the estimates generated from this sample could result. Coverage error is one of the most widely recognized shortcomings in Internet-based surveys if the goal is to draw inference from the results to the general population. In most countries around the world Internet access does not approach 100% penetration and the “haves” differ from the “have-nots” thus causing large coverage problems. As mentioned earlier the digital divide is associated with different distribution of socio-economic variables such as age, income, and education. As the Internet access increases, both coverage and differences between the covered and the non-covered will decrease, minimizing the risk of coverage bias.

When studying particular groups, for example universities or private companies where every student or employee (every unit in the target population) has been assigned an e-mail address known to the researcher, coverage bias is not a problem. However, results from such studies are harder to generalize to the general population. In Study II and III the target population consisted of Swedish women born between 1943 and 1962 residing in Uppsala Health Care Region. The frame consisted of the personal identification number with addresses held by the Swedish Tax Agency among the target population that responded to the study questionnaire in 1991-92. Thus the coverage bias could be substantial. Another way of defining the target population is Swedish women born between 1943 and 1962 residing in Uppsala Health Care Region that responded to the questionnaire in 1991/92. The coverage error would then be minor. The problem with either of these is the possibility to generalise. This is discussed below.

In Study IV the target population was Swedish women age 18-45 (2005) and the population frame was the identification number with addresses for women aged 18-45 held by the Swedish Tax Agency (accessed 28th of December 2004). Although people will have died or migrated between the latest update of the lists and the send out of the invitations in the study, the coverage error is thought to be minor.

**Sampling error** - Contacting every unit in the frame is often financially and administratively impossible. Contacting every Swede is not feasible so instead a sample of people is selected, and they are then the targets for measurement. A systematic exclusion of some of the members of the sampling frame that are not given the chance (or a reduced chance) of selection results in sampling bias if these differ from the sampled individuals. In web-based surveys one major problem is to construct a frame from which a random sample can be drawn, as a population register containing e-mail addresses does not exist. As mentioned earlier, this is one of the reasons for the use of non-probability sampling, which then makes any statistical inference difficult, if not impossible.

In Study II, III and IV a mail address list constituted the frame, and from there a random sample was drawn. All women in the sampling frame therefore had a known non-zero chance of being chosen, minimizing the risk of sampling error.

**Non-response bias** – This is the product of the non-response rate and the difference between the responders and the non-responders. As such the response rate alone is not a quality indicator. A high response rate is not a guarantee for the lack of non-response bias. It does, however, reduce the magnitude of non-response bias. As selective non-response might bias prevalence measures in epidemiological studies, the declining response rates are
a growing concern. This is a major problem in web surveys where the responders differ from non-responders (mentioned in coverage error) and the response rates are low, thus creating non-response bias.

In Study II and III, the information used in the analyses was collected in the original questionnaire in 1991-92 and thus included information for the non-responders of the 2003 questionnaire. We could therefore calculate the non-response bias. In Study IV we compared the responders of the different modes. Non-response bias was therefore not a problem.

### 8.1.2 Information bias

Information bias can occur when the information collected from or around study subjects is erroneous. The information collected can be misclassified, if for example a person is categorised as being a heavy smoker when he/she is, in fact, a light smoker. Misclassification can be either non-differential or differential. Non-differential misclassification i.e. the classification is not related to the outcome or the exposure, produces estimates biased towards the null. Differential misclassification, i.e., the classification is related to either the outcome or the exposure, can either exaggerate or underestimate the estimated association between outcome and exposure.

**Measurement error** – The discrepancy between the respondent mean response and the true sample mean. For example, the web-questionnaire layout and structure differ on different participants’ computers due to different technical prerequisites, such as processing power, connection speed and browsers used. The presentation of the questions and answers might therefore differ between participants creating measurement errors. This is one of the more easily addressed problems when using web-questionnaires. The knowledge and software exist so with tests, pilots and close communication between researchers and programmers, these problems can be solved. Also, the technical prerequisites, including skip patterns and plausibility and range checks help decrease measurement error.

In Study II-IV the questionnaires had been piloted and discovered discrepancies with regard to design and layout were corrected. In both cases firms with experience from the survey field and potential problems with layout and design of the questionnaire were employed, thus minimizing the risk of measurement error.

In Study II and III, different enterprises collected the paper and web questionnaires respectively, and we therefore do not believe misclassification of the outcome (web response vs. paper response) is a major concern. As for the exposure variables, these were self-reported in 1991-92 and could include some misclassification (erroneously entered answers), although this is likely to be non-differential. Alternatively, a correctly entered response in 1992 could be out of date in 2003, for example, by the increased level of education during the 11 years prior to the follow-up. One could hypothesize that the probability is higher that the younger women to a larger extent received additional education. This differential misclassification would most likely bias our results toward the null, underestimating the estimates. However, the youngest women were 29 years in 1991 and it is not likely that the proportion of women of this age group that increased their level of education is large enough to cause any substantial bias. Parity could likewise have biased the estimates towards the null, as the probability of having more children is higher for younger women. For the remaining socio-demographic variables, we do not believe there is reason to suspect the presence of a measurement error.
8.1.3 Confounding bias

A confounder is a factor associated with both the outcome and the exposure, which alters the estimated association between exposure and outcome. The confounding factor may however not be caused by the outcome or the exposure. A confounding factor can bias in either direction, towards the null or away from the null. If sufficient information about the confounding factor exists, it can be controlled for in the statistical analysis. In Study II-IV the presence of confounding factors was investigated by adjusting for such factors in the analysis. These included the exposure variables under study. As the changes between the crude and the adjusted estimates were minor, we do not believe confounding by these factors is a problem. There is also the possibility of the presence of other confounders that either are unknown to us or that we do not have enough information about. These could affect our results.

Caution should be taken when the results of Study II-IV are analyzed, with regard to the actual responses given. The mode effect, that is the effect the different modes may have on the responses, is a problem when using a mixed-mode model. This problem seems to be affected also by the topic of the survey: sensitive issues such as alcohol and illicit drug use might induce larger differences compared to issues regarding health care in general. There is no consensus for the mode effect when using web and paper modes and more research looking at different populations and different issues is needed.

Study I has not been mentioned in the discussions previously, due to its different study design. Study I is an observational study using a non-random selection of websites to be included in the analysis. The selection of websites investigated was not random as the aim was to replicate a search performed by a layperson. Furthermore, the quality criteria used in this study were chosen as they were clearly defined and easily measured giving little room for subjective evaluation and susceptibility to bias.

8.1.4 Generalisability

External validity refers to the extent to which the results of a study are generalisable to a different population. Designing web-studies necessitates conquering several methodological obstacles, the main being coverage error. A large part of the studies conducted using the web are performed in smaller populations with high Internet access using probability-sampled participants, such as university staff or students. Alternatively, larger samples of non-probability selected volunteers are used (for example Internet chat-room populations). Results generated from such populations cannot be generalised to a population that is not similar to the study population, i.e., does not have the same distribution of Internet access and socio-demographic variables associated with Internet use.

Due to the dynamic nature of the Internet, the generalisability of the results generated in Study I is limited. It should be seen as a snapshot of the contents available on the Internet at a given time period. However, the replication of the study a year later, as well as the more limited search in 2005 increases the probability that the picture captured is representative of the cancer risk sites available.
The prerequisites offered in Sweden with population-based registers and high Internet penetration makes it the optimal target population for studying web-based questionnaires. The generalisability of the results in Study II and III, however, is limited to a female population, age 39-60 with a similar level of Internet access and depending on the choice of target population, responders to a questionnaire 12 years prior to the current study. Whether the target population is all women invited in 1991/92 or all women responding in 1991/92, the possibility of generalising the results are limited. In the first example, the coverage bias causes low internal validity. In the second, the population to which the results are generalisable is a population of women that has already participated in a study. This latter problem is one occasionally seen in studies using a longitudinal design similar to the one used in our study.

The results generated in Study IV are generalisable to a female population, aged 18-45 with a comparable level of Internet access. Currently these populations might be sparse but the trend in Internet penetration around the world is positive. Hopefully it is just a matter of time before the results are generalisable to a larger number of female populations.

8.2 FINDINGS AND IMPLICATIONS

8.2.1 Study I

The purpose of Study I was to investigate the quality of cancer risk sites found on the Internet. The search was designed so as to mimic that of a layperson, and was performed in 2001 and 2002. An additional, more limited search was performed in 2005 to test for a change in trend.

The number of cancer risk sites increased by 50% between 2001 and 2002 (12 in 2001 and 18 in 2002). Only two out of 22 cancer risk sites fulfilled the quality criteria adequately. As such, our results are in line with those from previous studies.87-89 No signs of a change in trend (with regard to met quality criteria) were noted in January 2005. The quality of the information present on the websites found in both searches did not improve and there was no sign of a change in trend in 2005.

The results generated in Study I are, due to the design and the inherent construction of the Internet, not generalisable to Internet today. Nevertheless, it is an exceedingly burning topic. Despite the increase in Internet use as well as the growing amount of easily accessible health information on the Internet, there is no international consensus for quality criteria for the control and evaluation of information available online. This shortfall could not only lead to anxiety due to misinformation, it could also potentially hurt the relationship between the medical/scientific society and the general population. As mentioned, the wealth of the Internet empowers people, but the inability to differentiate between qualitative information and non-qualitative information might increase the sensation of mistrust. In order to strengthen the belief in the medical departments and health organizations these need to keep working so as to maintain compliance and trust. They have a responsibility to communicate their medical knowledge, not only within scientific circles, but also to the public, who is the present and future receiver of the medications and treatments generated from this research. If action is not taken, we will miss out on an exceptional opportunity to use the potential of the Internet – the possibility to reach beyond socio-demographic limitations and geographical boarders in an extremely
efficient way. It is important that organizations such as medical universities and governmental agencies assume their responsibility and provide trustworthy, high quality cancer information/tests on the Internet.

8.2.2 Study II

In Study II we wished to investigate the feasibility of using the web mode to collect health related data in a population-based study of middle-aged women in Sweden. A response rate of 72% out of which 41% were web responses implies that the web can be used successfully. Level of education, income and being a non-smoker were positive predictors of participation. A comparison between the web and paper responders did not show any large differences. Hence, our results support previous conclusions that the web can be used as a supplement for traditionally used methods. Another aim was to assess the potential differences between the women of the different response groups with regard to socio-demographic pattern. Since we had information about responders and non-responders, the potential bias in the whole study population as well as in the different groups (web and paper) could be studied. Even though differences with regard to level of education, income and smoking between the web responders and the paper responders exist, the potential bias caused by these is not large and what is more, it is not greater using the web compared to the paper mode. Thus far, this has been a major methodological concern and the survival of the method as a data collection tool in population-based epidemiological studies is, amongst other things dependent on whether or not selection bias can be avoided.

Our study, as well as other recently performed web-based studies, gives us reason to look with optimism on the future use of web-based methods in epidemiological studies. We conclude that web-based questionnaires are feasible in population-based epidemiological studies among women in Sweden. Although response rates were higher among non-smoking women with higher level of education and income, the potential bias caused is not different from the one caused by traditionally used methods.

8.2.3 Study III

In Study III our aim was to study the socio-demographic variable distribution between different groups of participants in a web-questionnaire including the lurkers, a group uniquely identified in web-studies. Contrary to our expectations the lurkers, subsequently non-responders bore closer resemblance to the completers (web) with regard to age and education than did the lurkers that subsequently completed a paper questionnaire. Surprising was also the fact that the lurkers differed significantly from each other. Prior to the advent of web-surveys, these groups were unidentifiable. More effort should be put into further characterization of these groups and to study the behavioural patterns and rational related to drop out. Given that they are tentatively willing to respond, lurkers are important targets for countermeasures aimed at increasing response rates.

As hypothesized, the degree of difficulty of questions in a web-questionnaire and the order in which they are presented affect drop out. This is in agreement with previous studies suggesting that layout and questionnaire length affect drop out. The “Easy-to-hard”
version of the web-questionnaire receives a 6 % unit higher response rate compared to the “Hard-to-easy” version. This represents a simple and efficient mean of increasing response rates, with no additional costs. Decreasing response rates have been noted during the last centuries, irrespective of mode used for data collection. Preventing this trend is vital for the validity of future epidemiological research. For paper surveys, the literature states several means to increase response rates, including pre-notification, reminders, incentives etc. However, these are expensive means of increasing the response rate and the increases found (4-16% units with pre-notification or reminders) are in the same range as the gains possible by the design alterations suggested in this study. More research is needed to study means of increasing response rates in web surveys.

8.2.4 Study IV

The aim of Study IV resembled that of Study II, namely the feasibility of using the web in a mixed-mode approach in a population-based cohort of 25,000 Swedish women. The differences were that the cohort in Study IV included younger age groups and that the women in this study had not been contacted previously.

There was no difference in the final response rate of women randomised to web mode, group A, versus web-and-paper responders, group B, (63%). We consider this response rate to be acceptable, particularly given the sensitive topic (sexual behaviours) of the study, which previously has been reported to generate lower response rates.

After the original send out, group A differed significantly from group B for all investigated socio-demographic variables except smoking. After two reminders, overall, students were more likely to respond to a web option whenever one was offered. Similarly, a higher level of education positively affected the risk of responding to web compared to paper mode. Age has been shown to negatively affect the risk of responding to web-questionnaires, which is contradicted by the results found in Study II. In the present study the only effect of age was to negatively affect the risk of responding to a web compared to a paper questionnaire in the first reminder group B.

The risk of item non-response was higher among paper responders compared to web responders both among general and sensitive questions. The differences were nonetheless smaller when looking at the sensitive questions. This could be the effect of a conscious choice; women deliberately skip the sensitive questions to a larger degree compared to the more general questions. The latter type of question might therefore have more to gain from the reminder function offered when using a web mode.

We conclude that this is a feasible mean to collect health data from a population-based cohort of Swedish women. Furthermore, we found no differences in socio-demographic variable distribution between the two groups of responders, something that might otherwise bias the estimates generated. The Internet use and access, although differing somewhat between the responders to web compared to paper, is high (90%) and does not solely explain the choice of response mode.
8.3 IMPLICATIONS FOR THE FUTURE

Today, a wealth of health information is constantly within reach on the Internet. As quality guarantees are hard to create and hard or even impossible to implement on the entire content of the Internet, ultimately it is up to the searcher of information to evaluate the information found. The discussion should be held in several parts of society on governmental, municipal, and regional levels. These organisations should aim to inform the public about qualitative-ensured websites, markers of quality, information about how to qualitatively evaluate a health site and the necessity to bear in mind to control the source of the information found.

It is our conviction that effectively used, the web mode could pave a new way in population-based epidemiological research, positively affecting response rates and minimizing drop out. The results presented in Study II and IV in this thesis are a proof-of-concept, confirming the feasibility of using the web in epidemiological studies. Study III concluded that web-based methods offer the possibility to characterise the lurkers, opening new possibilities for research on drop out and web-questionnaire construction, something not possible using traditional methods.

Nonetheless, we are in the early stages of this new field and much remains to be explored and empirically tested, particularly research from larger samples with available background information. Focus needs to be on deepening our knowledge regarding how to make use of this tool to the fullest of its capacities. Clearly, this should be investigated in light of the methodological constraints that exists - caution is justified. But at the same time it is vital that researchers keep an open mind whilst studying this new mode. The web is a technically advanced method with numerous advantages to be added to the existing modes of data collection. Through its technological innovations, interesting and advanced methodological research is made possible, for example real time randomization and rotation of questions, which in addition could reduce measurement error. The flexibility through several formats (e-mail with embedded survey, paper with URL to web site), the convenience of data entry and analysis, the low administration cost and ease of follow-up, facilitate the collection of large samples.

A major concern often cited by the adversaries of the web mode is non-response bias. However, this is a constant concern irrespective of method and might not be larger in web-studies, which was confirmed in Study II. It is also important to recall, when comparing the web and paper modes, that the paper mode is no gold standard. This method is not perfect and might not represent “the truth”. The view that the traditionally used methods are inherently superior, will compromise any effort to show the advantages and possibilities of the web mode.
9 CONCLUSIONS

- The quality of cancer risk sites on the Internet during the study period was poor and no improvement was seen in the follow-up searches. The searcher of health information needs to be made aware of the occurrence of health information of deficient quality.

- Web-based questionnaires are feasible in population-based epidemiological studies among women aged 18-60 in Sweden.

- Education, income, and being a student is positively associated with responding to a web-questionnaire. Besides these, there were no differences in the distribution of socio-demographic variables among the responders to web and paper-questionnaires. The bias associated with collecting information using web-questionnaires was not greater than that caused by paper-questionnaires.

- Given that they are tentatively willing to respond, lurkers, uniquely identified in web-based studies, present another important target for countermeasures aimed at increasing response rates. The level of difficulty of questions in a web-questionnaire and the order in which they are presented affect drop out and thus present another simple and efficient mean of increasing response rates.
10 SVENSK SAMMANFATTNING

Givet att de tekniska förutsättningarna finns tillhands kan vi, varhelst i välden vi än befinner oss, tillgodogöra oss internationella, nationella och lokala nyheter, betala räkningar, söka information, kommunicera med omvärlden – allt via Internet. Internet medför oanade möjligheter både till spridning och insamling av information och mängden tillgänglig information på Internet är enorm. Människor använder i allt större utsträckning Internet för att hitta information, inklusive hälsoinformation, både till sig själva och till sina nära och kära. Inom forskarsamhället är man sämre på att utnyttja de potentiella fördelarna med Internet. Samtidigt blir populationerna allt rörligare och geografiskt utspridda något som försvårar insamlandet av data för epidemiologiska studier.

Denna avhandling och de studier som ingår, faller inom ämnesområdet e-epidemiologi. Detta nya vetenskapliga fält omfattar insamling och användande av epidemiologisk kunskap och information med hjälp av digitala media, så som Internet, mobiltelefoner, digitalt papper och digital-TV. Vi har studerat tillgängliga cancertestsidor och kvaliteten på dessa samt insamlandet av hälsoinformation – i bågge fallen med hjälp av Internet.


I studie II-IV använder vi Internet i kombination med traditionella pappersenkäter för att samla in hälsoinformation från svenska kvinnor i två stora populationsbaserade undersökningar med 50 000 respektive 25 000 kvinnor, 18-60 år. Vi visade att det är möjligt att använda Internet för insamling av information i stora epidemiologiska studier. Utbildnings- och inkomstnivå var de faktorer som främst skiljde sig åt mellan de som svarade på Internet respektive de som svarade på pappersenkät. Trots dessa skillnader fann vi ingen eller likartad systematisk avvikelse från de samband som kan ses i hela den tillfrågade populationen för båda svarsmedlen. Detta är viktigt viktigt för att metoderna ska kunna användas inom den epidemiologiska forskningen. Ett annat viktigt resultat var att påvisa möjligheten att studera ”lurkers” (deltagare som loggar in på en webbstudie och börjar svara på frågorna i enkäten men som av någon anledning väljer att avbryta innan slutförrandet av enkäten), vilka endast går att karaktärisera och utforska då Internet används för informationsinsamling. Denna grupp är potentiella svarare och öppnar således för forskning kring metoder för att förhindra bortfall samt för att öka svarsfrekvensen.

Tack vare dagens tekniska förutsättningar krymper världen med avseende på avståndet mellan människor. Människor förflyttar sig med lättighet och populationen är följaktligen mer dynamiska. Internet är stor del i denna utveckling. Forskarsamhället har ett ansvar att bidra med kvalitativ information till dem som söker hälsoinformation på Internet. Likaså bör forskarsamhället inom epidemiologisk forskning utöka användandet av Internet som
tack vare sina unika egenskaper möjliggör gränsöverskridande och effektiv insamling av information från en allt rörligare population.
I would like to articulate my genuine appreciation to all of you who, in one way or another, contributed to the completion of this thesis.

**Jan-Eric Litton** – my main supervisor. I do not know how I got so lucky as to get such an excellent supervisor as well as a mentor. The novelty of the area made this trip exciting as well as unpredictable and if it were not for your exceptional guidance, brilliant mind, and faith in me, this thesis would never have existed. I am forever grateful.

**Paul Dickman** – my co-supervisor. With your superior teaching skills and never ending patience you guided me forwards and taught me to actually enjoy statistics. You made it fun!

**Elisabete Weiderpass** – my co-supervisor. Your expertise within the field of epidemiology made working with you a true pleasure.

**Per Hall** – the “goofy” professor with such an excellent sense of humor who brought me to MEB and co-authored my first article. I feel so lucky to have had the opportunity to work with such a brilliant scientist!

**Pär Sparén** – my co-author and exceptional epidemiologist that I am delighted to have had the chance to work with.

**Mats Lambe** – a splendid epidemiologist who introduced me to the field of epidemiology. Thank you!

**Nancy Pedersen** and **Hans-Olov Adami** – current and former head of department, thank you for creating and maintaining such a superior research environment.

**Cecilia Magnusson** – I would probably not have started my PhD if it were not for you (for which I will always blame you!) – you are a role model for female researchers and I am truly happy to have been given the opportunity to work with you.

**Åsa Klint** – for invaluable help with statistics and for uncountable chats about other important issues such as yoga and men!

**Christin Bexelius, Catherine Tuvblad, Linda Lindström, Lena George, Anna Svensson, Eva Dannetun** – thank you so much for friendship and invaluable scientific advice. Big thanks to all former and present PhD colleagues that made these four years endurable!

**Unnur Valdimarsdottir** – for your invaluable friendship and support – whether it be work or life in general.

**Pontus Lindqvist** – for great times - whether at MEB, at conferences or downing a beer (always such good fun!).
Gunilla Sonnebring and Helena Andersson for kindly proofreading my thesis and to all the guys in the IT-corridor, my corridor, for invaluable help with computers, software and for fun times.

For all the women in the Women’s Lifestyle and Health as well as the Women’s Health cohorts – without your participation this thesis would never have been possible.

Outside MEB

Caroline – my twin-sister. For always being there, for knowing me so well and for calmly advising me in the wisest direction.

Bodil (Dr Rawal) – for your honest and straightforward way of supporting me through four years of science and life. And for a lot of laughs!

Signe – for your contagious optimism, your never-ending encouragement and exceptional wit. I am so lucky to have you as a friend.

Agnes, Ebba, Maria – for endless support, such fun times, and everything else that true friends give.

Nila, Lisa – to my “hermanitas” – for always being there, just like sisters. Gabbis, Helena, Ika, Malin, Nina, Sofia, Antonia, Brenda and my other amazing friends that in one way or another have helped me through these four years. Your being there has made all the difference.

Agneta – my fairy godmother. Thank you for being there, listening and caring.

Pima – för att du är världens bästa farmor.

Mamma, Pappa, Oscar – I could never fully express how much you mean to me nor how much you have meant for the completion of this thesis. Oscar – your silent support and faith in me is invaluable - I am immensely proud to be your sister. Mami – soy la que soy gracias a tí. Pappa – my true hero and role model in life. This is all for you.
12 REFERENCES


33. Leece P, Bhandari M, Sprague S, Swiontkowski MF, Schemitsch EH, Tornetta P, Devereaux PJ, Guyatt GH. Internet versus mailed questionnaires: a randomized comparison (2). J Med Internet Res 2004;6:e30.


43. Bowling A. Mode of questionnaire administration can have serious effects on data quality. J Public Health (Oxf) 2005;27:281-91.


72. DISCERN.
86. Denscombe M. Web-Based Questionnaires and the Mode Effect: An Evaluation Based on Completion Rates and Data Contents of Near-Identical Questionnaires Delivered in Different Modes. Social Science Computer Review 2006;24:246-254.


