E-epidemiology: a comprehensive update

MMH van Gelder1*, A Pijpe2†

Abstract

Introduction

Interests in using information technology and specifically the Internet for epidemiological research (e-epidemiology) are increasing since it is expected to add several advantages to the traditional modes of data collection. These may include increased response rates, improved data quality and a more rapid and cost-efficient method for data collection and study population recruitment. In this study, we provide an overview of the latest developments in this relatively new field within epidemiologic research, focusing on methods of data collection and recruitment of study participants.

Conclusion

Information technology for data collection within epidemiological research offers many possibilities for large-scale studies: the current developments in the field of e-epidemiology show promising results, especially with regard to Web-based questionnaires, which showed that depending on the study population, the response rates for Web-based questionnaires may be higher than for paper-based questionnaires. However, future research should focus on the internal and external validity of the data collected, as well as on other aspects of e-epidemiologic research, including participant recruitment through the Internet, use of Web 2.0 content, data collection through smartphone applications, short message service and videoconferencing, and inherent privacy and ethical concerns.

Discussion

The authors have referenced some of their own studies in this review. These referenced studies have been conducted in accordance with the Declaration of Helsinki (1964) and the protocols of these studies have been approved by the relevant ethics committees related to the institution in which they were performed. All human subjects, in these referenced studies, gave informed consent to participate in these studies.

Web-based questionnaires

In populations with high Internet access rates, Web-based questionnaires may be considered as an alternative for or complementary method of data collection to the traditional modes. The advantages of Web-based questionnaires are straightforward: (1) data quality is improved through validation checks during questionnaire completion, automated skipping of non-applicable items and not having to make the error-prone steps of data entry and coding, (2) automatic data management systems may be used to send invitations for follow-up questionnaires and reminders by e-mail, (3) information is returned more rapidly compared with paper-based questionnaires and (4) the costs of data collection may be reduced. Indeed, some large-scale cohort studies, such as the Snart Gravid study, the PRIDE Study and the Nightingale Study, use Web-based questionnaires as their primary method of data collection.

However, one of the major concerns of Web-based questionnaires is bias resulting from survey non-response, which leads to selection bias if the association between exposure and outcome differs between

Licensee OA Publishing London 2013. Creative Commons Attribution License (CC-BY)

participants and all targeted subjects. A 2008 meta-analysis showed that mail surveys had higher response rates than Web surveys, but studies conducted in the last five years that directly compared response rates for Web-based and paper-based questionnaires showed that the gap in response rates is closing (Table 1)–14, especially in higher educated populations and when e-mail invitations were used. According to the same meta-analysis, postal invitations to complete a Web-based questionnaire may negatively influence response rates due to the additional step of switching from the postal invitation to the Internet by typing in the web address, login and password for participation, which could explain up to 15% lower response rates for Web-based questionnaires, and thereby also the lower response rates for Web-based questionnaires compared to paper-based questionnaires in the studies by Boschman et al.7 and Sinclair et al.10.

Another concern regarding Web-based questionnaires is the validity of the data collected. Compared with the traditional modes of data collection, Web-based questionnaires were expected to yield larger measurement errors due to simple errors, including faster reading by Internet users, the respondents not scrolling down to all questions or answering options and suboptimal questionnaire design.15,16. Although evidence on the validity of data collected through Web-based questionnaires is still limited (Table 2), the validity of data on anthropometric, dietary and socio-demographic factors is generally reported to be high.17–26. More studies to validate data collected through Web-based questionnaires are needed to establish whether this approach is suitable for epidemiological research in various study populations. However, it should be noticed that none of the traditional methods is perfect either.

### Table 1 Summary of studies conducted since 2008 that directly compared response rates for Web-based and paper-based questionnaires

<table>
<thead>
<tr>
<th>Study population</th>
<th>Age (years)</th>
<th>Study year</th>
<th>Recruitment details</th>
<th>Response rate WBQ</th>
<th>Response rate PBQ</th>
<th>Difference</th>
<th>P-value</th>
<th>References</th>
</tr>
</thead>
<tbody>
<tr>
<td>Bricklayers</td>
<td>46 (17–64)*</td>
<td>2009</td>
<td>Postal invitation</td>
<td>28.4% (42/148)</td>
<td>43.5% (64/147)</td>
<td>−15.1%</td>
<td>0.007</td>
<td>Boschman et al.7</td>
</tr>
<tr>
<td>Cancer patients</td>
<td>59 (30–80)†</td>
<td>2007–2008</td>
<td>NR</td>
<td>63.9% (39/61)</td>
<td>63.9% (39/61)</td>
<td>0.0%</td>
<td>1.00</td>
<td>Kamo et al.8</td>
</tr>
<tr>
<td>Construction supervisors</td>
<td>49 (19–65)*</td>
<td>2009</td>
<td>Postal invitation</td>
<td>43.0% (64/149)</td>
<td>45.9% (67/146)</td>
<td>−2.9%</td>
<td>0.61</td>
<td>Boschman et al.7</td>
</tr>
<tr>
<td>General populationd</td>
<td>18–80†</td>
<td>2010</td>
<td>E-mail/postal invitation</td>
<td>93.4% (170/182)</td>
<td>89.1% (188/211)</td>
<td>+4.3%</td>
<td>0.14</td>
<td>Lagerros et al.9</td>
</tr>
<tr>
<td>Households</td>
<td>NR</td>
<td>2011</td>
<td>Personalised postal invitation</td>
<td>4.7% (289/6210)</td>
<td>10.5% (575/5500)</td>
<td>−5.8%</td>
<td>&lt;0.001</td>
<td>Sinclair et al.10</td>
</tr>
<tr>
<td>Households</td>
<td>NR</td>
<td>2011</td>
<td>Generic postal invitation</td>
<td>2.2% (101/4515)</td>
<td>7.5% (299/4000)</td>
<td>−5.3%</td>
<td>&lt;0.001</td>
<td>Sinclair et al.10</td>
</tr>
<tr>
<td>Members AEA</td>
<td>NR</td>
<td>NR</td>
<td>E-mail/postal invitation</td>
<td>52.5% (672/1281)</td>
<td>42.0% (538/1280)</td>
<td>+10.5%</td>
<td>&lt;0.001</td>
<td>Greenlaw and Brown-Welty11</td>
</tr>
<tr>
<td>Patients chiropractor clinic</td>
<td>18+*</td>
<td>2008</td>
<td>In-clinic administration</td>
<td>74.6% (47/63)</td>
<td>62.8% (135/215)</td>
<td>+11.8%</td>
<td>0.08</td>
<td>Derby et al.12</td>
</tr>
<tr>
<td>PhD holders</td>
<td>43†</td>
<td>2008</td>
<td>Personalised e-mail/postal invitation</td>
<td>64.8% (250/386)</td>
<td>48.8% (62/127)</td>
<td>+16.0%</td>
<td>0.001</td>
<td>Barrios et al.13</td>
</tr>
<tr>
<td>University students</td>
<td>24†</td>
<td>NR</td>
<td>E-mail/postal invitation</td>
<td>54.4% (536/986)</td>
<td>50.7% (498/983)</td>
<td>+3.7%</td>
<td>0.10</td>
<td>Whitehead14</td>
</tr>
<tr>
<td>Total</td>
<td></td>
<td></td>
<td></td>
<td>15.8% (2210/13981)</td>
<td>19.5% (2465/12670)</td>
<td>−3.7%</td>
<td>&lt;0.001</td>
<td></td>
</tr>
<tr>
<td>Total excl.</td>
<td></td>
<td></td>
<td></td>
<td>55.9% (1820/3256)</td>
<td>50.2% (1591/3170)</td>
<td>+5.7%</td>
<td>&lt;0.001</td>
<td></td>
</tr>
</tbody>
</table>

AEAs, American Evaluation Association; NR, not reported; PBQ, paper-based questionnaire; WBQ, web-based questionnaire.
*response rate WBQ minus response rate PBQ; †median (range); ‡mean (range); †among those subjects that agreed to participate; †range; †mean.

Recruitment of study participants through the internet

Besides approaching potential study participants through a postal letter or e-mail, the Internet is increasingly being used for study recruitment purposes. The first experiences with the social networking site Facebook as a method to recruit a specific target group for health-related research showed modest success. Facebook allows advertisement targeting based on age, gender, location and listed interests, which makes it possible to show the advertisement only to users that potentially fulfil the study eligibility criteria. To evaluate the advertisement’s performance, Facebook provides a number of metrics, including the number of times the advertisement is shown to a user (‘impressions’) and the number of clicks the advertisement received. The costs are determined either as cost per 1000 impressions or cost per click. For health-related studies, the click-through rates varied between 0.02% and 0.05%27,28, and the final costs of the Facebook campaign per completed survey ranged between £3.25 for a study recruiting young US cigarette users (n = 1548)28 and £25 for a study recruiting Italian pregnant women (n = 8)29. Similarly, but less often reported in the scientific literature, Google advertising (AdWords) can be used. Using this service, a short advertisement is displayed when a user searches for certain keywords, of which the settings can be adjusted based on the budget, targeted locations, time of the day and day of the week. For a study recruiting participants for an online preventive depression intervention, the click-through rate was 6.01%, resulting in an average cost of $10.75 per signed-up person (n = 602)30. Other methods for online recruitment could include posts in forums, Internet support groups and noticeboards, respondent-driven sampling (snowballing), link exchanges with relevant web sites and sending e-mails to specific groups or lists.

In 2009, Edwards et al.31 published a systematic review in which they identified multiple strategies to increase response to Web-based questionnaires. In addition to multiple design issues such as shorter questionnaires, using a white background and simple header and personalised invitations, non-monetary incentives can be used to increase response rates. In Web-based

Licensee OA Publishing London 2013. Creative Commons Attribution License (CC-BY)


Table 2 Results of studies validating data on anthropometric, dietary and socio-demographic factors collected through Web-based questionnaires

<table>
<thead>
<tr>
<th>Variable/instrument</th>
<th>Reference standard</th>
<th>Results</th>
<th>References</th>
</tr>
</thead>
<tbody>
<tr>
<td>Activity energy expenditure (seven-day recall)</td>
<td>Doubly labelled water</td>
<td>r = 0.14 (p = 0.54)</td>
<td>Namba et al.17</td>
</tr>
<tr>
<td>Activity energy expenditure (diary)</td>
<td>Doubly labelled water</td>
<td>r = 0.68 (p &lt; 0.001)</td>
<td>Namba et al.17</td>
</tr>
<tr>
<td>Alcohol consumption</td>
<td>Telephone interview</td>
<td>kappa = 0.88</td>
<td>Graham et al.18</td>
</tr>
<tr>
<td>Anthropometry</td>
<td>Paper-based questionnaire</td>
<td>ICC = 0.86–1.00; kappa = 0.7–0.9</td>
<td>Touvier et al.19</td>
</tr>
<tr>
<td>BMI (weight, height)</td>
<td>Anthropometric measurements</td>
<td>Se = 0.69; Sp = 1.00</td>
<td>Bae et al.20</td>
</tr>
<tr>
<td>Body weight</td>
<td>In-home assessment</td>
<td>r = 0.98a</td>
<td>West et al.21</td>
</tr>
<tr>
<td>Dietary assessment measures</td>
<td>Paper-based questionnaire</td>
<td>r = 0.28–0.73</td>
<td>Beasley et al.22</td>
</tr>
<tr>
<td>Dietary assessment measures</td>
<td>Four-day food record</td>
<td>r = 0.21–0.64</td>
<td>Beasley et al.22</td>
</tr>
<tr>
<td>Dietary assessment measures</td>
<td>Twenty-four-hour recall</td>
<td>r = 0.10–0.57</td>
<td>Beasley et al.22</td>
</tr>
<tr>
<td>General health status (RAND-36)</td>
<td>Paper-based questionnaire</td>
<td>ICC = 0.54–0.75</td>
<td>Bliven et al.23</td>
</tr>
<tr>
<td>Perceived health status</td>
<td>Telephone interview</td>
<td>kappa = 0.73</td>
<td>Graham et al.18</td>
</tr>
<tr>
<td>Socio-demographic factors</td>
<td>Paper-based questionnaire</td>
<td>ICC = 0.81–1.00; kappa = 0.76–1.00</td>
<td>Vergnaud et al.24</td>
</tr>
<tr>
<td>Total energy expenditure</td>
<td>Doubly labelled water</td>
<td>r = 0.45</td>
<td>Arab et al.25</td>
</tr>
<tr>
<td>Total energy expenditure (seven-day recall)</td>
<td>Doubly labelled water</td>
<td>r = 0.59 (p = 0.006)</td>
<td>Namba et al.17</td>
</tr>
<tr>
<td>Total energy expenditure (diary)</td>
<td>Doubly labelled water</td>
<td>r = 0.88 (p &lt; 0.001)</td>
<td>Namba et al.17</td>
</tr>
<tr>
<td>Total energy expenditure (Active-Q)</td>
<td>Doubly labelled water</td>
<td>r = 0.52 (p &lt; 0.001)</td>
<td>Bonn et al.26</td>
</tr>
<tr>
<td>Total household income</td>
<td>Telephone interview</td>
<td>kappa = 0.93</td>
<td>Graham et al.18</td>
</tr>
</tbody>
</table>

BMI, body mass index; ICC, intraclass correlation coefficient; r, correlation coefficient; Se, sensitivity; Sp, specificity.
aalthough highly correlated, self-reported weight was on an average approximately 1 kg lower compared with the weight measured by a trained professional.
studies, lotteries are popular incentives to increase response and completion rates. However, a few recent studies on their effectiveness show that the effects are small. In an online panel, lotteries may slightly increase response among those panelists with a low motivation to respond (e.g. if the subject did not participate in a previous study). For community-based surveys, a lottery with a small number of large prizes was the most cost-effective incentive compared with no incentive, prepaid cash, which yielded the highest response rate, and a lottery with a relatively large number of small prizes.

User-driven Internet content: Web 2.0
Data resulting from participation in Web 2.0, which refers to web sites that allow users to interact and communicate with each other to produce user-generated content, may yield new possibilities for epidemiologic data collection. Participation in Web 2.0 takes place on three levels. Firstly, it can occur unnoticed to the user, such as by using a search engine that keeps track of the searches performed (even stratified by geographic location and time, for example Google Trends) or by clicking on web feeds. Secondly, it may include active searching of user-generated content, for example reading hotel or restaurant reviews. Finally, it comprises the production of content to be read by others, for example content in blogs, microblogs (Twitter), social networking sites (Facebook, LinkedIn) and photo-sharing sites (Instagram, Flickr, Pinterest).

The best-known example of using Web 2.0 content for epidemiologic research is the study by Ginsberg et al., who used Google search query data to track influenza-like illness in the United States. The relative frequency of selected queries was highly correlated with the data from the traditional surveillance networks and the resulting influenza-like illness estimates were 1–2 weeks ahead of the surveillance network reports. Recently, Twitter data were used for the same purpose, in which sophisticated methods to distinguish between actual flu cases and people merely talking about it were applied. Furthermore, Twitter may produce other data valuable for epidemiological research, including measuring behavioural risk factors, symptoms and medication use, but it is as yet unknown what new empirical knowledge can be gained by using this method of data collection. User profiles of social networking sites could also contain a wealth of data on risk behaviours, relationships and attitudes, whereas blog posts can automatically be screened for more detailed information on personality, health status and psychological profiles. Moreover, user-driven Internet content often holds detailed geographical information, either through automatically publishing the user’s location concurrently with a status update or through use of specialized applications, such as Foursquare, Google Latitude and Find My Friends, which could be of interest for epidemiologists quantifying mobility patterns in relation to environmental exposures, activities and health outcomes.

For clinical epidemiologists, social networking sites that focus on health-related problems may be of special interest. For example, on PatientsLikeMe, almost 190,000 people with various conditions share data on many factors, such as their disease status, symptoms, treatments, quality of life and weight. With information from these social networking sites, a number of research studies have already been conducted. The PatientsLikeMe web site has been integrated with ClinicalTrials.gov, so that patients can search for trials they might be eligible for. DailyStrength has over 500 support groups and health blogs; in which, users share their medical conditions, mental health issues and life challenges. On MedHelp, patients can document their health by using condition-specific applications and personal health records, which can be tracked by other members (and researchers). Furthermore, companies that offer direct-to-consumer genetic testing, such as 23andMe, create databases with genetic profiles, which can be used for research purposes. For example, two novel loci for Parkinson’s disease were identified using data from customers of 23andMe.

Other methods for data collection using information technology
In addition to Web-based questionnaires and Web 2.0, some other methods for data collection using information technology have been explored on a small scale for application in epidemiological studies. If mobile phone numbers of potential study participants are known, short message service (SMS) may be used to administer very short, easy questionnaires with one question for each SMS and set answering options. In two Swedish studies using SMS to assess influenza vaccination and physical activity in subjects randomly sampled from the population registry, 45% and 91% of subjects who received the questions completed the questionnaires, respectively. In these studies, SMS data on the variables of interest were comparable with data originating from more traditional methods of data collection. Another application of SMS is that it may be used to send reminders for completing Web-based questionnaires.

Smartphones may be more feasible to collect epidemiological data than low-end mobile phones since they combine the functions of a personal digital assistant, portable media players, cameras and global positioning system navigation and allow the user to install and run dedicated applications (‘apps’). The first

Licensee OA Publishing London 2013. Creative Commons Attribution License (CC-BY)

studies showed promising results for collecting data on physical activity (using the device’s accelerometer and gyroscope or an app)⁴⁴,⁴⁵, infant feeding practices⁴⁶ and health-related diaries⁴⁶. Future research should develop and make these tools available to the full extent of their capabilities for health-related research, in which new possibilities, such as real-time exposure measurement and instant individual feedback, could be explored.

Videoconferencing has been proposed as a low-cost alternative for face-to-face interviews, especially in studies in which study participants are geographically dispersed over a large area, provided that both a high-bandwidth connection (i.e. ≥ 1024 kbps) and the hardware and software are both present⁴⁴. However, a German randomised pilot study among young adults showed that interviewing using Skype was not feasible: the response rate was 10% (95% CI: 5%–15%) in the Skype group and 22% (95% CI: 15%–28%) in the telephone interview group⁴⁵.

Concerns

New developments in epidemiological research, such as using information technology for medical research, generate new discussions, even though some of them refer to the basic epidemiological principles of internal and external validity. Although Internet access rates are high in developed countries, still a selective part of the population uses all the possibilities the Internet has to offer, which may be a threat to the generalisability of the results of Internet-based studies. Comparable with the traditional paper-based questionnaires and personal and telephone interviews, response rates for Web-based questionnaires are less than 100%. Mixed-mode or crossover designs may be used to increase survey participation rates. Data on response rates for studies using SMS, smartphone apps and videoconferencing are insufficient and need further investigation before they can be added to the epidemiological toolbox. Reassuringly, the traditional methods of data collection have shown little bias in exposure-outcome associations resulting from selective participation⁴⁶,⁴⁷, and this is expected not to be different for e-epidemiological tools. Regarding user-driven Internet content, there is not only self-selection in participants in Web 2.0, but there is also selection in the information available, as the user decides which information he shares on the Internet. Furthermore, even for Web-based questionnaires, information on data quality is often lacking and specifically for Web 2.0 content, data on potential confounding factors are often unavailable, possibly leading to residual confounding.

In addition to the discussions about the internal and external validity of the results of e-epidemiological studies, privacy and ethical issues are being raised, especially concerning studies using Web 2.0 content⁴⁸. Users of Web 2.0 are likely unaware that the information they share may be used for research purposes, even though this is stated in the web site’s User Agreement. Questions may be raised whether clicking on the ‘I agree’ button for the web site’s terms of service represents an acceptable informed consent for health-related research, especially for web sites that offer direct-to-consumer genetic testing. However, the use of Web-based questionnaires also raises questions since there are no general guidelines regarding participant protection: how should informed consent be obtained? Is an online-administered form sufficient or should it be a paper-based, signed form? Which security measures should be taken to protect the (personal) data collected? How can we verify the identity of the participants? Institutional Review Boards and other regulatory bodies may need to develop guidelines to solve these issues.

References


Conclusion

Using information technology and specifically the Internet for data collection and participant recruitment within epidemiological research offers many possibilities for large-scale studies. Web-based questionnaires are increasingly being considered as a fully-fledged alternative to the traditional modes of data collection with, in many populations, response rates comparable to those for paper-based questionnaires. However, many issues need to be further investigated and tested before e-epidemiology can be completely implemented in health-related research. These include determining the internal and external validity of the data of e-epidemiological studies, addressing privacy and ethical concerns regarding participant recruitment through the Internet and use of Web 2.0 content, and the development of data collection through SMS, smartphone apps or videoconferencing. Network meetings and collaborations between epidemiologists pioneering in this relatively new research area, such as the 2012 e-epi symposium in Cardiff and the initiation of focus group e-Epidemiology of The Netherlands Epidemiology Society, certainly contribute to further improvements of e-epidemiological tools.
Critical review


36. Using Twitter to track the flu: researchers find a better way to screen the tweets. http://releases.jhu.edu/2011/01/24/twitter-flu/.

Critical review