Journal of Medical Internet Research

Impact Factor (2018): 4.945 - ranked #1 medical informatics journal by Impact Factor
Volume 8 (2006), Issue 4 ISSN: 1438-8871 Editor in Chief: Gunther Eysenbach, MD, MPH

Contents

Original Papers

Original Paper

Age-Specific Search Strategies for Medline

Monika Kastner[1,2], MSc; Nancy L Wilczynski[2], MSc; Cindy Walker-Dilks[2], MLS; Kathleen Ann McKibbon[2], MLS, PhD; Brian Haynes[2,3,*], MD, PhD

[1]Department of Health Policy Management and Evaluation, Faculty of Medicine, University of Toronto, Toronto, ON, Canada

[2]Health Information Research Unit, Department of Clinical Epidemiology and Biostatistics, Faculty of Health Sciences, McMaster University, Hamilton, ON, Canada

[3]Department of Medicine, Faculty of Health Sciences, McMaster University, Hamilton, ON, Canada

[*]for the Hedges Team

Corresponding Author:
Brian Haynes, MD, PhD
McMaster University
Hamilton, ON L8N 3Z5
Canada
Phone: +1 905 525 9140 ext 24931
Fax: +1 905 577 0017
Email: bhaynes@mcmaster.ca

Abstract

Background: Many clinicians and researchers are interested in patients of a specific age (childhood, geriatrics, and so on). Searching for age-specific publications in large bibliographic databases such as Medline is problematic because of inconsistencies in indexing, overlapping age categories, and the spread of the relevant literature over many journals. To our knowledge, no empirically tested age-specific search strategies exist for Medline.

Objective: We sought to determine the retrieval characteristics of age-specific terms in Medline for identifying studies relevant for five clinical specialties: adult medicine, geriatric medicine, pediatric medicine, neonatal medicine, and obstetrics.

Methods: We compared age-specific search terms and phrases for the retrieval of citations in Medline with a manual hand search of the literature for 161 core health care journals. Six experienced research assistants who were trained and intensively calibrated read all issues of 161 journals for the publishing year 2000. In addition to classifying all articles for purpose and quality, study participants' ages were also recorded. Outcome measures were sensitivity, specificity, precision, and accuracy of single and combination search terms.

Results: When maximizing sensitivity, the best sensitivity and specificity achieved with combination terms were 98% and 81.2%, respectively, for pediatric medicine, 96.4% and 55.9% for geriatric medicine, 95.3% and 83.6% for neonatal medicine, 94.9% and 64.5% for adult medicine, and 82% and 97.1% for obstetrics. When specificity was maximized, all disciplines had an expected decrease in sensitivity and an increase in precision. Highest values for optimizing best sensitivity and specificity were achieved in neonatal medicine, 92.5% and 92.6%, respectively.

Conclusion: Selected single terms and combinations of MeSH terms and textwords can reliably retrieve age-specific studies cited in Medline.

(J Med Internet Res 2006;8(4):e25) doi:10.2196/jmir.8.4.e25

KEYWORDS

Information storage and retrieval; Medline; medical subject headings

Introduction

Clinicians and researchers seeking research reports for specific age categories, including generalists and those who are engaged in clinical specialties such as adult medicine, geriatric medicine, pediatric medicine, neonatal medicine, or obstetrics, need to target their literature searches so that the information they retrieve is relevant to their patient population. Difficulty in finding pertinent evidence contributes to the challenges health professionals have in keeping up-to-date and practising evidence-based medicine [1-7].

Finding age-specific evidence in Medline is a difficult task for several key reasons. In large bibliographic databases such as Medline, optimal search retrieval for individual topics is

hampered by the overwhelming amount of available information that is not pertinent to the question. When users search in Medline they have the potential to retrieve articles from any of the approximately 4800 journals that are currently indexed in the database. The size of this general purpose biomedical database coupled with imperfections in indexing [1-3] lead to a high risk of missing articles that are relevant to the topic of the search while at the same time retrieving many articles that are off target. Effective ways to refine the search may be helpful for those wanting to keep up-to-date and for those looking for an answer to a specific patient care question.

Searching in Medline for a specific patient population by selecting "age-specific" journals will not help because studies relevant to any age group are scattered through a wide range of journals, including general journals that cater to no particular age group. Moreover, in Medline, the indexing practices used to identify the ages of those involved in a study are so liberal that they create a very imprecise representation of the age categories of the participants within the study. Medline indexers apply all relevant age-specific index terms to an article regardless of how many participants fall within that category. Thus, if just one patient or participant in the study falls into a particular age category, that age-specific medical subject heading (MeSH) term will be applied. For example, if a researcher was interested in intercultural communication in family medicine around issues of newborn care, the study by Harmsen and colleagues [8] might be retrieved using the following index terms: infant, newborn; ethnic groups; communication; and family practice. However, looking at the patient population studied, only 0.9% of the participants were children under the age of 12 years—likely very few of these were newborns. The study included participants from many age categories, resulting in eight age-specific index terms being assigned to this article (infant, newborn; infant; child, preschool; child; adolescent; adult; middle aged; and aged). For searchers who are interested in communication around newborn issues, this article is likely not useful even though the indexing indicates that it is potentially relevant. These age-classification problems are compounded by the less than optimal search strategies used by clinicians, including their lack of knowledge about how to narrow searches without missing relevant information, and their uncertainty about when to stop searching [9,10].

To assist clinicians searching for studies on age-specific patient populations, we have developed and tested Medline search strategies for detecting studies for specific age categories as well as tested age-specific search terms pertinent to five age-related clinical specialties. In this paper, we report on the evaluation of the retrieval performance of age-specific search strategies in Medline compared with a manual review (the "gold standard" search) of each article in every issue of 161 journals in the year 2000.

Search strategies are useful tools when searching in large electronic databases. We previously developed search strategies for use in Medline to detect clinically relevant scientifically sound articles in the areas of causation, prognosis, treatment, and diagnosis [11-15]. After publishing our initial work on search strategy development [15], we were approached by neonatologists and gerontologists to develop age-specific search strategies because they expressed frustration with the inefficiency of the current system for finding content specific to their patient population. Using only the age-related MeSH terms when searching can be time-consuming because retrievals can be very large and imprecise. To our knowledge, no empirically developed age-specific search strategies have been previously reported for Medline.

Methods

The study compared the retrieval performance of age-specific search terms and phrases in Medline (accessed using Ovid) with a manual review of each article in every issue of 161 journal titles for the year 2000. The 161 journals were chosen over several years in an iterative process based on a hand search review of over 400 journals. The journals were recommended by clinicians, librarians, editors, and publishers and were chosen based on Science Citation Index impact factors and ongoing assessment of their yield of studies and reviews of scientific merit and clinical relevance [16] in the production of 4 evidence-based medicine secondary journals (*ACP Journal Club*, *Evidence-Based Medicine*, *Evidence-Based Nursing*, and *Evidence-Based Mental Health*). The 161 journals include content for the disciplines of internal medicine (eg, *Annals of Internal Medicine*), general medical practice (eg, *BMJ*, *JAMA*, and *Lancet*), mental health (eg, *Archives of General Psychiatry*, *British Journal of Psychiatry*), and general nursing practice (eg, *Nursing Research*).

Six research assistants hand searched the 161 journals for the year 2000 and collected data on age of the study participants according to our hand search categories defined in Table 1. This data collection was part of a larger study in which the research assistants applied methodological criteria to each article in each issue to determine if the article was methodologically sound for seven purpose categories (eg, treatment and diagnosis). All purpose category definitions and corresponding methodological rigor have been outlined in previous papers [4,17]. Research staff were rigorously calibrated for applying all these criteria, including the age classification of study participants, and interrater agreement for application of all criteria exceeded 80% beyond chance $\kappa = 0.81$; 95% CI = 0.79-0.84) [4].

Table 1. Comparison of hand searching and Medline MeSH classification of age categories

Hand Search Category	Our Definition	Medline MeSH Term Category	MeSH Definition
Fetus	Fetus	-	-
Newborn	Birth to 1 month	Infant, newborn	Birth to 1 month
Infant	> 1 month to < 24 months	Infant	1 to 23 months
Preschool	2 years to < 6 years	Child, preschool	2 to 5 years
Child	6 years to < 13 years	Child	6 to 12 years
Adolescent	13 years to < 19 years	Adolescent	13 to 18 years
Adult	19 years to < 45 years	Adult	19 to 44 years
Middle age	45 years to < 65 years	Middle aged	45 to 64 years
Aged	65 years to < 80 years	Aged	65 to 79 years
Aged 80	≥ 80 years	Aged, 80 and over	80 years and over
ND	Nondiscernible	-	-

MeSH terms and textwords related to age (eg, infant, child, adult) were downloaded from Medline and were treated as "diagnostic tests" for detecting studies with an age-specific population as determined by a hand search of the literature from 161 journals (the gold standard). The hand search data were obtained by reading each issue completely. The downloaded Medline data from the 161 journals included the retrieval sets for each of the individual terms. After these two data sources were obtained (ie, the Medline downloads and the hand search review), a database was created that included the matched merged content from these two sources. These Ovid retrieval sets were then manipulated by our own set of programs to calculate our outcome measures—the operating characteristics of each age-specific searching term (eg, sensitivities, specificities, and precision) for individual terms and for combinations of terms. When we merged the two data sets (Medline and hand search), we determined the match. If Medline included an item that was not indexed, we went back to the journal and scored it. If we had scored an item that was not in Medline, we removed it from the merged database. Therefore, the final merged database included only items that had hand search scores and Medline indexing. This merged database was used to develop the age-specific search strategies [17].

Borrowing from the concepts of diagnostic test evaluation and library science, we determined the sensitivity, specificity, precision, and accuracy of single- and multiple-term Medline searches. We considered these operating characteristics as the indicators of search term performance. Sensitivity for a given age-specific topic is defined as the proportion of relevant articles (ie, articles with the desired age-specific content) that are retrieved; specificity is the proportion of nonrelevant articles (ie, articles that are outside the desired age-specific content) not retrieved; precision is the proportion of retrieved articles that are relevant (a library science term that is equivalent to "positive predictive value" in diagnostic test evaluation); and accuracy is the proportion of all articles that are correctly classified (ie, overall proportion of relevant articles retrieved and nonrelevant articles not retrieved). Our hand search of the 161 journals indexed in Medline led to the classification of all articles in these journals for age-related content. Search terms were then tested to determine their performance in retrieving age-relevant articles while eliminating those that were nonrelevant. An automated process (which we developed and implemented using a computer program) was used to calculate the operating characteristics (performance) for each single and combination term in Medline. Formulae for calculating the operating characteristics (ie, sensitivity, specificity, precision, and accuracy) of searches are shown in Table 2.

Table 2. Formulae for calculating the sensitivity, specificity, precision, and accuracy of searches for detecting age-specific articles*

	Hand Search	
Detection of Search Terms	**Meets Criteria**	**Does Not Meet Criteria**
Detected	a	b
Not detected	c	d
	a+c	b+d

*Sensitivity = a/(a + c); precision = a/(a + b); specificity = d/(b + d); accuracy = (a + d)/(a +b + c + d). All articles classified during the manual review of the literature, n = (a + b + c + d).

Individual search terms with sensitivity > 25% and specificity > 75% for a given age category were incorporated into the development of search strategies that included two or more terms. All combinations of terms used the Boolean "OR." For the development of multiple-term search strategies to either optimize sensitivity or specificity, we tested all two-term search strategies with sensitivity of at least 75% and specificity at least 50%.

To construct a comprehensive set of search terms, a list of MeSH terms and textwords was initially generated, and input was sought from clinicians and librarians in the United States and Canada through interviews with known searchers, requests at meetings and conferences, and requests to the National Library of Medicine. These experts were asked which terms or phrases they used when searching for age-specific studies, as well when searching for studies in specific purpose categories. Search terms could be MeSH terms, including publication types and subheadings, or textwords specific to age in titles and abstracts of articles. Various truncations were also applied to the textwords, phrases, and MeSH terms. We compiled a list of 543 age-specific terms (Multimedia Appendix). All terms were tested in Medline using the Ovid Technologies searching system.

Age categories for the hand search were modeled from the MeSH terms used to index age content. A comparison of hand search categories and MeSH term definitions is shown in Table 1. The major difference between the hand search age categories and the MeSH terms is in how they were applied. During the hand search, we classified the age of study participants in primary studies or review articles in the following way: select one age category, if possible, or up to three to represent where ≥ 50% of participants fell. This procedure is intended to more accurately represent the focus of age-category research of clinical relevance than the comprehensive indexing of all participants' ages provided by the Medline index terms (which may be more pertinent for nonclinical purposes).

We defined five age-specific specialty areas by collapsing our hand search age categories (see Table 1) and through discussions with clinicians from each specialty area about which definition most appropriately reflected the age of their patients in clinical practice: geriatric medicine (≥ 65 years of age), adult medicine (19 to < 65 years of age), pediatric medicine (> 1 month to < 19 years of age), neonatal medicine (birth to 1 month), and obstetrics (fetus).

Results

Tables 3 to 7 show the operating characteristics of top-performing combinations of terms with best sensitivity, best specificity, and best optimization of sensitivity and specificity while minimizing the difference between the two, for detecting studies on geriatric medicine, adult medicine, pediatric medicine, neonatal medicine, and obstetrics in Medline in 2000. Search strategies are reported using Ovid's search engine syntax for Medline (mp = multiple posting—term appears in title, abstract, or subject heading; sh = subject heading [MeSH term]; tw = textword—word or phrase appears in title or abstract; : = truncation; pt = publication type; exp = explode—a search term that automatically includes closely related MeSH terms; tu = therapeutic use as a subheading; xs = exploded subheading).

Geriatric Medicine

The single term "exp adult" yielded the best sensitivity (96.4%) with a specificity of 55.9% for retrieving articles about geriatric medicine. However, by using the next best sensitivity combination, "aged.sh. OR age:.tw.", a small sacrifice in sensitivity (1% absolute decrease) resulted in a much better specificity compared with the most sensitive term (absolute increase 14.4%) and improved precision (absolute increase 5.2%) and accuracy (absolute increase 13.3%). As expected, precision improved slightly when specificity was maximized (absolute increase 8.6%). The term that yielded the best optimization of sensitivity and specificity, "aged.sh.", resulted in 93.6% sensitivity and 82.7% specificity.

Table 3. Combination of terms with the best sensitivity, best specificity, and best optimization of sensitivity and specificity for detecting studies about geriatric medicine (≥ 65 years) in Medline in 2000

Search Strategy*	Operating Characteristics†			
	Sensitivity, % (95% CI) (n = 3309)	Specificity, % (95% CI) (n = 45719)	Precision, %‡ (95% CI)	Accuracy, % (95% CI) (n = 49028)
Best sensitivity	96.4	55.9	13.7	58.7
(exp adult)	(95.8-97.1)	(55.5-56.4)	(13.2-14.1)	(58.2-59.1)
Next best sensitivity	95.4	70.3	18.9	72.0
(aged.sh. OR age:.tw.)	(94.7-96.1)	(69.8-70.7)	(18.2-19.4)	(71.6-72.3)
Best specificity	63.3	84.0	22.3	82.6
(aged, 80 and over.sh. OR of age.tw.)	(61.7-65.0)	(83.7-84.4)	(21.5-23.1)	(82.3-83.0)
Next best specificity	93.6	82.7	28.2	83.5
(aged.sh.)	(92.8-94.5)	(82.4-83.1)	(27.3-29.0)	(83.1-83.8)
Best optimization of sensitivity and specificity	93.6	82.7	28.2	83.5
(aged.sh.)	(92.8-94.5)	(82.4-83.1)	(27.3-29.0)	(83.1-83.8)

*Search strategies are reported using Ovid's search engine syntax for Medline (if a single search term is shown, this term outperformed two- and three-term combinations). Best sensitivity while keeping specificity ≥ 50%; Best specificity while keeping sensitivity ≥ 50%; Best Optimization of Sensitivity and Specificity is based on lowest possible absolute difference between sensitivity and specificity; exp = explode, a search term that automatically includes closely related indexing terms; sh = subject heading; : = truncation; tw = textword (word or phrase appears in title or abstract).

†Total database has 49028 articles, of which 3309 articles are relevant to geriatric medicine and 45719 are irrelevant to geriatric medicine.

‡n varies by row.

Adult Medicine

The three-term strategy "adult.mp. OR middle aged.sh. OR age:.tw." yielded the best sensitivity (94.9%) and had a specificity of 64.5% for retrieving articles about adult medicine. When specificity was maximized (85.2%) with the single term "middle aged.sh.", sensitivity lowered to 72.3%, but precision improved to 62.1% (absolute increase 14.8%) and accuracy improved as well (absolute increase 9.8%). The best optimization of sensitivity and specificity occurred with the combined terms "middle aged.sh. OR of age.tw.", with values approaching 79%.

Table 4. Combination of terms with the best sensitivity, best specificity, and best optimization of sensitivity and specificity for detecting studies about adult medicine (19 to < 65 years) in Medline in 2000

Search Strategy*	Operating Characteristics †			
	Sensitivity, % (95% CI) (n = 12307)	Specificity, % (95% CI) (n = 39721)	Precision, %‡ (95% CI)	Accuracy, % (95% CI) (n = 49028)
Best sensitivity	94.9	64.5	47.3	72.1
(adult.mp. OR middle aged.sh. OR age:.tw.)	(94.5-95.3)	(64.4-64.9)	(46.6-47.8)	(71.7-72.5)
Best specificity	72.3	85.2	62.1	81.9
(middle aged.sh.)	(71.5-73.1)	(84.8-85.5)	(61.3-62.8)	(81.6-82.3)
Next best specificity	75.3	81.4	57.6	80.0
(adult.sh.)	(74.6-76.1)	(81.0-81.8)	(56.8-58.4)	(79.5-80.3)
Best optimization of sensitivity and specificity	78.7	77.9	54.4	78.1
(middle aged.sh. OR of age.tw.)	(78.0-79.4)	(77.4-78.3)	(53.7-55.1)	(77.7-78.5)

*Search strategies are reported using Ovid's search engine syntax for Medline (if a single search term is shown, this term outperformed two- and three-term combinations). Best sensitivity while keeping specificity ≥ 50%; Best specificity while keeping sensitivity ≥ 50%; Best Optimization of Sensitivity and Specificity is based on lowest possible absolute difference between sensitivity and specificity; mp = multiple posting—term appears in title, abstract, or subject heading; sh = subject heading; : = truncation; tw = textword (word or phrase appears in title or abstract).

†Total database has 49028 articles, of which 12307 articles are relevant to adult medicine and 39721 are irrelevant to adult medicine.

‡n varies by row.

Pediatric Medicine

The three-term strategy "child:.mp. OR adolescent.mp. OR infan:.mp." yielded the best sensitivity of 98.0% with a specificity of 81.2% for retrieving articles about pediatric medicine. When specificity was maximized (97.1%) with the single term "children.tw.", a striking trade-off in sensitivity occurred as it was lowered to 58.2% (absolute decrease 39.8%). Yet, as expected, precision improved (absolute increase 30.9%). The three-term strategy "adolescent.tw. OR children.tw. OR child, preschool.sh." yielded the best optimization of sensitivity and specificity (89.3% and 87.3%, respectively).

Table 5. Combination of terms with the best sensitivity, best specificity, and best optimization of sensitivity and specificity for detecting studies about pediatric medicine (> 1 month to < 19 years) in Medline in 2000

Search Strategy*	Operating Characteristics†			
	Sensitivity, % (95% CI) (n = 2845)	Specificity, % (95% CI) (n = 46183)	Precision, %‡ (95% CI)	Accuracy, % (95% CI) (n = 49028)
Best sensitivity (child:.mp. OR adolescent.mp. OR infan:.mp.)	98.0 (97.4-98.5)	81.2 (81.1-81.4)	24.6 (23.8-25.4)	82.4 (82.1-82.8)
Best specificity (children.tw.)	58.2 (56.4-60.0)	97.1 (97.0-97.3)	55.5 (53.7-57.2)	94.9 (94.7-95.1)
Best optimization of sensitivity and specificity (adolescent.tw. OR children.tw. OR child, preschool.sh.)	89.3 (88.1-90.4)	87.3 (87.0-87.6)	30.3 (29.3-31.3)	87.4 (87.1-87.7)

*Search strategies are reported using Ovid's search engine syntax for Medline (if a single search term is shown, this term outperformed two- and three-term combinations). Best sensitivity while keeping specificity ≥ 50%; Best specificity while keeping sensitivity ≥ 50%; Best Optimization of Sensitivity and Specificity is based on lowest possible absolute difference between sensitivity and specificity; mp = multiple posting—term appears in title, abstract, or subject heading; : = truncation; tw = textword (word or phrase appears in title or abstract); sh = subject heading.

†Total database has 49028 articles, of which 2845 articles are relevant to pediatric medicine and 46183 are irrelevant to pediatric medicine.

‡n varies by row.

Neonatal Medicine

Best sensitivity (95.3%) was achieved by the three-term strategy "infan:.mp. OR child:.mp. OR gestation:.tw.", with a specificity of 83.6% for retrieving articles about neonatal medicine. An expected trade-off occurred in sensitivity (absolute decrease 41.7%) with the most specific term, "infants.tw." (98.7%). However, precision increased to 38.2% (absolute increase 30.8%) and accuracy reached 98.2%. The three-term strategy "infan:.mp. OR gestation:.tw. OR neonatal.tw." yielded the best optimization of sensitivity and specificity, reaching values of 93% (which were the highest among all five specialties).

Table 6. Combination of terms with the best sensitivity, best specificity, and best optimization of sensitivity and specificity for detecting studies about neonatal medicine (birth to 1 month) in Medline in 2000

Search Strategy*	Operating Characteristics†			
	Sensitivity, % (95% CI) (n = 663)	Specificity, % (95% CI) (n = 48365)	Precision, %‡ (95% CI)	Accuracy, % (95% CI) (n = 49028)
Best sensitivity	95.3	83.6	7.4	83.8
(infan:.mp. OR child:.mp. OR gestation:.tw.)	(93.7-96.9)	(83.3-83.9)	(6.8-7.9)	(83.4-84.1)
Best specificity	53.6	98.7	38.2	98.2
(infants.tw.)	(52.6-60.2)	(98.6-98.8)	(34.9-41.0)	(98.0-98.3)
Next best specificity	67.7	98.2	33.7	97.8
(infants.tw. OR neonatal.tw.)	(64.0-71.1)	(98.0-98.3)	(31.0-36.0)	(97.6-97.9)
Best optimization of sensitivity and specificity	92.5	92.6	14.7	92.6
(infan:.mp. OR gestation:.tw. OR neonatal.tw.)	(90.5-94.5)	(92.4-92.8)	(13.6-15.7)	(92.4-92.8)

*Search strategies are reported using Ovid's search engine syntax for Medline (if a single search term is shown, this term outperformed two- and three-term combinations). Best sensitivity while keeping specificity ≥ 50%; Best specificity while keeping sensitivity ≥ 50%; Best Optimization of Sensitivity and Specificity is based on lowest possible absolute difference between sensitivity and specificity; mp = multiple posting—term appears in title, abstract, or subject heading; : = truncation; tw = textword (word or phrase appears in title or abstract).

†Total database has 49028 articles, of which 663 articles are relevant to neonatal medicine and 48365 are irrelevant to neonatal medicine.

‡n varies by row.

Obstetrics

The combination of terms "gestation:.tw. OR fetal.tw. OR pregnancy.tw." produced the best sensitivity of 82.0%, with a very high specificity of 97.1% for retrieving articles about obstetrics. The maximization of specificity (reaching almost 99%) with the single term "gestation:.tw." yielded a 1.8% increase in specificity but with a marked trade-off in sensitivity, which decreased to 52.0% (absolute decrease 30%).

Table 7. Combination of terms with the best sensitivity, best specificity, and best optimization of sensitivity and specificity for detecting studies about obstetrics (fetus) in Medline in 2000

Search Strategy*	Operating Characteristics†			
	Sensitivity, % (95% CI) (n = 516)	Specificity, % (95% CI) (n = 48512)	Precision, %‡ (95% CI)	Accuracy, % (95% CI) (n = 49028)
Best sensitivity	82.0	97.1	23.4	97.0
(gestation:.tw. OR fetal.tw. OR pregnancy.tw.)	(78.7-85.3)	(97.0-97.3)	(21.4-25.3)	(96.9-97.1)
Best specificity	52.0	98.9	33.6	98.4
(gestation:.tw.)	(47.6-56.3)	(98.8-99.0)	(30.2-36.7)	(98.3-98.5)
Best optimization of sensitivity and specificity	80.7	79.3	4.0	79.3
(pregnancy.tw. OR fetal.tw. OR age:.tw.)	(77.2-84.0)	(78.9-79.7)	(3.6-4.4)	(79.0-79.7)

*Search strategies are reported using Ovid's search engine syntax for Medline (if a single search term is shown, this term outperformed two- and three-term combinations). Best sensitivity while keeping specificity ≥ 50%; Best specificity while keeping sensitivity ≥ 50%; Best Optimization of Sensitivity and Specificity is based on lowest possible absolute difference between sensitivity and specificity; : = truncation; tw = textword (word or phrase appears in title or abstract).

†Total database has 49028 articles, of which 516 articles are relevant to obstetrics and 48512 are irrelevant to obstetrics.

‡n varies by row.

Discussion

Our study shows that selected age-specific search strategies can achieve high retrieval of studies for age-specific populations. Our age-specific search strategies performed differently among the five specialties we investigated. The highest sensitivity and specificity were achieved for pediatric medicine (98% and 81.2%, respectively) and neonatal medicine (95.3% and 83.6%, respectively). This finding may be a result of these age groups being more precisely defined and that studies tend to be narrowly focused on them. Search strategies within obstetrics yielded a higher specificity (97.1%) than sensitivity (82%), indicating that this strategy was better at filtering out nonrelevant

age-specific articles than retrieving them. The best performing strategy for optimizing sensitivity and specificity was achieved within neonatal medicine (92.5% and 92.6%, respectively). In all cases, precision was low, a consequence of searching in large multi-purpose databases. Future research is focusing on potential ways to improve precision without compromising sensitivity, for example, by searching in journal subsets.

A possible limitation to our study is the generalizability of our findings to other publication years as our data was collected in the year 2000. We believe, however, that our search strategies are robust because no major changes have been made to age-specific MeSH terms since the year 2000. Moreover, we have previously shown that search strategies developed in 1990 were robust when searching in 2000 [18]. Another potential limitation of our study is that our interrater agreement for classifying age content did not reach 100%. However, exceeding the level of agreement achieved in our study (> 80% beyond chance) is rarely done in diagnostic studies. The scope of journals investigated in our journal subset could be a limitation, but we have no indication that these search strategies would perform differently in other journal subsets aside from the precision values reported. Precision is affected by the prevalence of on-target articles within the database. Thus, our precision figures are presented as estimates of search strategy performance.

The utility of age-specific filters will vary according to the needs of clinicians and researchers who must weigh the consequences of using a sensitive or specific search. Although a sensitive search will not miss many relevant articles, such searches are less precise and entail time-consuming sorting through irrelevant articles. The narrower yield of a specific search will capture many relevant articles and take less weeding, but it has greater potential for missing key articles.

Search Examples

To illustrate the use of age-specific search strategies, if a geriatrician was looking for information about current treatment strategies for Huntington disease, she might begin her search by entering the content term "Huntington disease" in Medline, which would yield 5907 articles (Table 8).

Table 8. Example: best sensitivity (keeping specificity ≥ 50%) search strategies for detecting treatment studies in geriatric medicine (patients ≥ 65 years of age) in Medline (1996 to July week 3, 2005)

Search Strategy*					
Content Term	**Boolean Operator**	**Best Sensitivity Combination Strategy for Treatment Studies**	**Boolean Operator**	**Best Sensitivity Combination Strategy for Geriatric Medicine**	**Number of Articles**
Huntington disease	-	-	-	-	5907
Huntington disease	AND	clinical trial.mp. OR clinical trial.pt. OR random:.mp. OR tu.xs.	-	-	901
Huntington disease	AND	clinical trial.mp. OR clinical trial.pt. OR random:.mp. OR tu.xs.	AND	exp adult†	483

*Search strategies are reported using Ovid's search engine syntax for Medline. mp = multiple posting—term appears in title, abstract, or subject heading; : = truncation; pt = publication type; tu = therapeutic use as a subheading; xs = exploded subheading; exp = explode—a search term that automatically includes closely related indexing terms.

†Outperformed two- and three-term combinations.

However, sifting through such a large number of articles would be time-consuming and many of these articles would not be relevant to treatment studies in geriatric medicine. By combining the content term "Huntington disease" with the most sensitive combination of terms for treatment studies (clinical trial.mp. OR clinical trial.pt. OR random:.mp. OR tu.xs.), the search can be narrowed to 901 articles. Further, by adding the most sensitive strategy for geriatric medicine (exp adult) to this search string with the Boolean operator AND, the search is refined to 483 articles, which is much more manageable than the original 5907 articles retrieved from searching the content term only. A sensitive search such as this would be an efficient beginning for researchers interested in conducting systematic reviews.

A more specific approach may be especially useful for physicians who do not have time to process an exhaustive search. In the above example, by combining the content word "Huntington disease" with the most specific search strategy for treatment studies [12], "randomized controlled trial.mp. OR randomized controlled trial.pt.", and the most specific search strategy for geriatric medicine, "aged, 80 and over.sh. OR age.tw.", the search yields five articles (Table 9). This is a dramatic reduction in the number of articles retrieved by searching the content term alone (5907 articles), but key articles can be missed.

Table 9. Example: best specificity (keeping sensitivity ≥ 50%) search strategies for detecting treatment studies in geriatric medicine (patients ≥ 65 years of age) in Medline (1996 to July week 3, 2005)

Search Strategy*					
Content Term	**Boolean Operator**	**Best Specificity Combination Strategy for Treatment Studies**	**Boolean Operator**	**Best Specificity Combination Strategy for Geriatric Medicine**	**Number of Articles**
Huntington disease	-	-	-	-	5907
Huntington disease	AND	Randomized controlled trial.mp. OR randomized controlled trial.pt.	-	-	46
Huntington disease	AND	Randomized controlled trial.mp. OR randomized controlled trial.pt.	AND	aged, 80 and over.sh. OR of age.tw.	5

*Search strategies are reported using Ovid's search engine syntax for Medline. mp = multiple posting—term appears in title, abstract, or subject heading; pt = publication type; sh = subject heading; tw = textword (word or phrase appears in title or abstract).

Conclusion

Selected age-specific search strategies can enhance the retrieval of studies for clinicians and researchers who need information relevant for a well-defined age-category patient population. The optimal trade-off between sensitivity and specificity should be determined according to the needs of the searcher.

Acknowledgments

This research was funded by the National Library of Medicine, USA(Grant no. 5 R01 LM06866-02). The Hedges Team includes Angela Eady, Brian Haynes, Susan Marks, Ann McKibbon, Doug Morgan, Cindy Walker-Dilks, Stephen Walter, Nancy Wilczynski, and Sharon Wong.

Conflicts of Interest

None declared.

Multimedia Appendix 1

Age terms used for developing age-specific search strategies [XLS (MS Excel) file, 48 KB - jmir_v8i4e25_app1.xls]

References

1. Haynes RB, Sackett DL, Tugwell P. Problems in the handling of clinical and research evidence by medical practitioners. Arch Intern Med 1983 Oct;143(10):1971-1975. [Medline: 84022898] [doi: 10.1001/archinte.143.10.1971]
2. Bennett NL, Casebeer LL, Kristofco RE, Strasser SM. Physicians' Internet information-seeking behaviors. J Contin Educ Health Prof 2004 Dec;24(1):31-38. [Medline: 23242417] [doi: 10.1002/chp.1340240106]
3. Lee KP, Schotland M, Bacchetti P, Bero LA. Association of journal quality indicators with methodological quality of clinical research articles. JAMA 2002 Jun 5;287(21):2805-2808 [FREE Full text] [Medline: 22035764] [doi: 10.1001/jama.287.21.2805]
4. Wilczynski NL, McKibbon KA, Haynes RB. Enhancing retrieval of best evidence for health care from bibliographic databases: calibration of the hand search of the literature. Medinfo 2001;10(Pt 1):390-393. [Medline: 21490557]
5. Haynes RB, Guyatt GH, Tugwell P. Clinical epidemiology: a basic science for clinical medicine. 2nd ed. Boston, MA: Little, Brown; 1991.
6. ; Evidence-Based Medicine. A New Approach to Teaching the Practice of Medicine. Evidence-Based Medicine Working Group. JAMA 1992 Nov 4;268(17):2420-2425. [Medline: 1404801] [doi: 10.1001/jama.268.17.2420]
7. McKibbon KA, Walker-Dilks CJ. The quality and impact of MEDLINE searches performed by end users. Health Libr Rev 1995 Sep;12(3):191-200. [Medline: 96359344] [doi: 10.1046/j.1365-2532.1995.1230191.x]
8. Harmsen H, Bernsen R, Meeuwesen L, Thomas S, Dorrenboom G, Pinto D, et al. The effect of educational intervention on intercultural communication: results of a randomised controlled trial. Br J Gen Pract 2005 May;55(514):343-350 [FREE Full text] [Medline: 15904552]
9. Ely JW, Osheroff JA, Ebell MH, Chambliss ML, Vinson DC, Stevermer JJ, et al. Obstacles to answering doctors' questions about patient care with evidence: qualitative study. BMJ 2002 Mar 23;324(7339):710 [FREE Full text] [Medline: 21906460] [doi: 10.1136/bmj.324.7339.710]
10. Balas EA, Stockham MG, Mitchell JA, Sievert ME, Ewigman BG, Boren SA. In search of controlled evidence for health care quality improvement. J Med Syst 1997 Feb;21(1):21-32. [Medline: 97316095] [doi: 10.1023/A:1022887224126]

11. Wilczynski NL, Haynes RB, Hedges Team. Developing optimal search strategies for detecting clinically sound causation studies in MEDLINE. AMIA Annu Symp Proc 2003:719-723 [FREE Full text] [Medline: 14728267]
12. Wilczynski NL, Haynes RB, Hedges Team. Developing optimal search strategies for detecting clinically sound prognostic studies in MEDLINE: an analytic survey. BMC Med 2004 Jun 9;2(1):23 [FREE Full text] [Medline: 15189561] [doi: 10.1186/1741-7015-2-23]
13. Haynes RB, McKibbon KA, Wilczynski NL, Walter SD, Werre SR, Hedges Team. Optimal search strategies for retrieving scientifically strong studies of treatment from Medline: analytical survey. BMJ 2005 May 21;330(7501):1179 [FREE Full text] [Medline: 15894554] [doi: 10.1136/bmj.38446.498542.8F]
14. Haynes RB, Wilczynski NL. Optimal search strategies for retrieving scientifically strong studies of diagnosis from Medline: analytical survey. BMJ 2004 May 1;328(7447):1040 [FREE Full text] [Medline: 15073027] [doi: 10.1136/bmj.38068.557998.EE]
15. Haynes RB, Wilczynski N, McKibbon KA, Walker CJ, Sinclair JC. Developing optimal search strategies for detecting clinically sound studies in MEDLINE. J Am Med Inform Assoc 1994 Nov;1(6):447-458 [FREE Full text] [Medline: 95153434]
16. McKibbon KA, Wilczynski NL, Haynes RB. What do evidence-based secondary journals tell us about the publication of clinically important articles in primary healthcare journals? BMC Med 2004 Sep 6;2(1):33 [FREE Full text] [Medline: 15350200] [doi: 10.1186/1741-7015-2-33]
17. Wilczynski NL, Morgan D, Haynes RB, Hedges Team. An overview of the design and methods for retrieving high-quality studies for clinical care. BMC Med Inform Decis Mak 2005;5(1):20 [FREE Full text] [Medline: 15969765] [doi: 10.1186/1472-6947-5-20]
18. Wilczynski NL, Haynes RB, Hedges Team. Robustness of empirical search strategies for clinical content in MEDLINE. Proc AMIA Symp 2002:904-908. [Medline: 22352856]

Abbreviations

MeSH: medical subject heading

submitted 21.12.05; peer-reviewed by R Leipzig, M van der Paardt; comments to author 20.03.06; revised version received 12.08.06; accepted 23.08.06; published 25.10.06.

Please cite as:
Kastner M, Wilczynski NL, Walker-Dilks C, McKibbon KA, Haynes B
Age-Specific Search Strategies for Medline
J Med Internet Res 2006;8(4):e25
URL: http://www.jmir.org/2006/4/e25/
doi:10.2196/jmir.8.4.e25
PMID:

Original Paper

Web Portals in Primary Care: An Evaluation of Patient Readiness and Willingness to Pay for Online Services

Kenneth G Adler[1], MD, MMM

Department of Family and Community Medicine, University of Arizona, Tucson, AZ, USA

Corresponding Author:
Kenneth G Adler, MD, MMM
Department of Family and Community Medicine
University of Arizona
5300 E Erickson, Suite 108
Tucson, AZ 85712
USA
Phone: +1 520 721 5330
Fax: +1 520 547 5743
Email: kadler@azacp.com

Abstract

Background: Online Web communication between physician and patient has been proposed by leading primary care organizations as a way to enhance physician-patient communication, but lack of payment for this service has acted as a significant barrier to implementation.

Objective: This study evaluates current patient readiness and willingness to pay for online services in a fairly typical urban family medicine practice.

Methods: All patients that visited the author for medical care during a one-month period in the spring of 2006 were anonymously surveyed with a one-page survey instrument that inquired about demographics, willingness to pay a small annual fee for online services, the greatest fee they were willing to pay, and their most desired service.

Results: A total of 346 patients out of 2380 active patients in the study practice (14.5%) were surveyed. The valid survey response rate was 95.1% (329/346.) Three quarters, or 75.4%, of patients had Internet access. The group with the highest access were 18- to 29-year-olds (97%), and the group with the least access were those 70 years and up (56%) ($P < .001$). Categorized by employment, students and employed patients had the best access at 92% and 87%, respectively, and retirees and disabled patients had the worst access at 66% and 42%, respectively ($P < .001$). Of all patients with Internet access, 74.6% (n = 185) were willing to pay a small annual fee for one or more of the following online services: viewing of parts of their medical record, messaging with their physician, medication refills, appointment requests, and billing inquiries. Willingness to pay did not vary significantly by age ($P = .06$). Of all respondents, regardless of Internet access, 47.1% (n = 155) were willing to pay US $10 or more per year, with the median amount being US $20. Of those with Internet access (n = 248), 60.1% (n = 149) were willing to pay US $10 or more per year, and 31% were willing to pay US $50 or more per year. The three most important services to patients with Internet access (n = 248), in order of importance, were emailing with their physician (34%), Internet viewing of their medical record (22%), and medication refills (11%) ($P < .001$).

Conclusions: This study suggests that patients of all ages are currently ready and willing to pay a small annual fee for online services with their primary care physician's office. If 47.1% of a practice of 2500 patients each paid US $10 per year for online services, the annual revenue generated would be US $11775. Not only does this study support the economic feasibility of patient Web portals, but it suggests that online services could form a new line of revenue for primary care physicians.

(J Med Internet Res 2006;8(4):e26) doi:10.2196/jmir.8.4.e26

KEYWORDS

Internet; communication; primary health care; electronic mail; patient access to records; Web-based services; payment schemes

Introduction

We live in a time when online communication has become commonplace in numerous service industries, yet to date that has not been the case in health care – at least as far as doctor-patient communication goes. Based on a paucity of research, there is a perception that patients want online communication with their physicians and their offices, but aren't willing to pay for it. A Harris Interactive survey of over 2000 online adults in 2002 showed that almost all (90%) respondents would like to communicate with their physicians online. This same survey showed that only 37% were willing to pay for it [1]. A prior Harris Interactive survey had shown that online adults are often unhappy with their ability to communicate with their physician and his/her office, and the majority felt that online access would improve communication [2]. A randomized controlled trial by Lin et al showed that patients in an academic internal medicine practice who used a Web portal had higher satisfaction with physician-patient communication than those who did not [3]. The American Academy of Family Physicians, the American College of Physicians, and the Society of General Internal Medicine have all proposed online communication as one tool to help revitalize primary care in the United States and to help improve doctor-patient communication and patient access [4-6].

Despite this generally acknowledged desire to implement widespread physician-patient online communication, a number of barriers exist, including lack of reimbursement, concerns about patient privacy and confidentiality, medicolegal concerns, practical workflow concerns, and physician fears of being overwhelmed by online messages [7].

Physicians fear that they will be inundated with online patient messages, and they voice frustration with not being reimbursed for these services [8]. Yet several studies of early adopters have not found physicians to be overwhelmed by patient emails [9-12]. The development of confidential Web portals that are linked to an electronic health record (EHR) have addressed many of the privacy, confidentiality, and workflow concerns, but lack of reimbursement remains a major obstacle. In one survey of physician users of Web messaging, 80% of physicians said they would be more willing to engage in online patient communication if it was reimbursed [12]. A few insurance companies have started paying physicians for direct online communication between doctor and patient for care of a clinical problem (e-consult) at a rate of about US $25 to $30, or half that of an intermediate office visit, but they remain the exception [13-15].

Patient Web portals represent a significant advance beyond traditional physician-patient email. They are online sites that can be free-standing or integrated with an EHR. Often they are tied to a personal health record (PHR), or, in the most recent definition of PHRs, Web portals are actually synonymous with PHRs as long as they include the ability to record and update key aspects of a person's health history, like their medication list, allergies, and health problem list. These sites allow patients to securely message their physician and to request medication refills and/or appointments. In some cases, when linked to an EHR, they allow patients to view and/or download some components of their medical record, such as medication and problem lists [16-18]. They offer convenient asynchronous communication for patient and physician alike [9,10]. Increased use of Web portals can reduce physician phone traffic and increase practice efficiency [19].

Despite all their advantages and the expressed desire of patients and primary care physician organizations to utilize Web portals, a number of questions about reimbursement remain. Do patients want online access to their physician office enough to pay a small annual fee for it, perhaps in addition to an e-consult transaction fee? Such a subscription fee might more than defray the direct and indirect costs of offering these services. It is axiomatic that a person's willingness to pay for services indicates a higher level of true demand than their simply indicating they would use the service if it were free.

This study was undertaken to determine the true level of demand for online services in a fairly typical family medicine practice in Tucson, Arizona and to answer the following related questions: What are the demographics of the patients most interested and least interested in these services, and what is their current Internet connectivity? Are patients willing to pay a nominal annual fee for these services and, if so, how much? Which of these services do patients value the most?

Methods

Practice Site

The practice used was the author's. This practice is part of a family medicine office that has 3 physicians and a nurse practitioner. Patients have a clearly identified physician and rarely, if ever, see one of the physician partners. The nurse practitioner primarily assists each physician with acute care patient visits and patient coverage when a physician is out of the office. In other words, the study practice, although part of a group practice, functions like a solo practice with a one-third time nurse practitioner. The office uses an EHR and has offered free "one-way" email to patients since 2000. If patients sign an email agreement, they are permitted to email the practice messages and the practice responds by phone. This service has never been heavily utilized. Based on EHR data, the practice size is 2380 active patients, defined as patients seen by the author in the last 36 months who are still alive. This practice sees virtually all insurances available in the community, including self-pay and Medicaid (called AHCCCS in Arizona.) It includes the full range of socioeconomic groups and ages (newborn to over 100 years). Being an urban family practice that does not involve obstetrics, it is likely more heavily weighted toward geriatric patients than those in a smaller or rural community. As a mature practice, it is largely closed to new patients except for family members of current patients.

Survey

The author's patients were given a one-page survey entitled "Web Portal Survey" (Multimedia Appendix) when they arrived for an appointment with the author from April 10, 2006 to May 11, 2006. A receptionist handed out the survey. The survey did not have any specific patient identifiers other than gender. Age

and employment were only generally identified by broad categories. Surveys were collected by reception when the patients checked out. They were given to the author in a nonsequential manner at the end of each day. The author did not discuss the content of the self-explanatory surveys with the patients other than to occasionally inquire if they had completed one. Patients were only allowed to complete one survey. If they returned for a follow-up visit during the study period (as determined by reviewing the schedule), they did not get another survey to complete. Parents completed surveys on behalf of patients younger than 18 years, and caregivers occasionally completed surveys for elderly or disabled patients who were unable to complete them themselves.

The one-page survey consisted of strictly check box answers and was easy to complete. The questions asked about simple demographics, information on Internet access, and willingness to pay a "small fee" for any of five different Web portal services. Quantification of the fee was obtained by inquiring the maximum annual fee respondents were willing to pay, and this was followed by asking which of the five services were most important to the respondent.

Given the manner in which the data were collected, handled, and entered into the database, the data were, by deliberate design, de-identified. This was done to honor the statement at the top of the survey that this was an anonymous survey, thus encouraging patients to be frank in their responses. The study was deemed to be exempt by the Human Research Committee (Institutional Review Board) of Tucson Medical Center.

Analysis

After collecting the surveys, the data were entered nonsequentially into an Access database (Microsoft, USA) and analyzed using select queries. When comparing respondent groups for significant differences, contingency tables were created and the StatsDirect statistical software v2.5.7 (Chesire, UK) was used. Analysis was done with the Fisher exact test. *P* values less than .05 were considered significant.

Results

Return Rate

During the study period, 346 unique patients were seen. Of those, 337 completed and returned a survey (return rate of 97.4%). A total of 8 surveys were invalidated due to missing data (not answering whether they had Internet access and/or not making any responses to the question on willingness to pay a small fee for one or more of the five online services listed on the survey). Thus, the number of valid surveys was 329 (response rate of 95.1%).

Internet Access

In the study practice, 75.4% of patients had Internet access. Internet access varied significantly by age, gender and employment (Table 1). Specifically, in terms of age, 18- to 29-year-olds had the highest access (97%), and patients 70 years and older had the least access (56%) ($P < .001$; Fisher exact test, 8x2 contingency table). Of note, 41% of the respondents were 60 and older.Students and employed patients had the best access, 92% and 87%, respectively, and retirees and disabled patients had the worst access, 66% and 42%, respectively ($P < .001$; Fisher exact test, 6x2 contingency table). Males were more likely to have Internet access than females ($P = .02$; Fisher exact test, 3x2 contingency table).

Table 1. Demographics and Internet access

	All Patients (N = 329)	Patients With Internet (N = 248)
	n (%)	n (%)
Gender		
Male	135 (41)	111 (45)
Female	192 (58)	136 (55)
Unknown	2 (1)	1 (0)
Age		
Under 18	14 (4)	12 (5)
18–29	29 (9)	28 (11)
30–39	26 (8)	21 (8)
40–49	50 (15)	44 (18)
50–59	63 (19)	47 (19)
60–69	61 (19)	46 (19)
70 and up	72 (22)	40 (16)
Unknown	14 (4)	10 (4)
Employment		
Student	26 (8)	24 (10)
Employed	142 (43)	123 (50)
Unemployed	12 (4)	8 (3)
Disabled	24 (7)	10 (4)
Retired	123 (37)	81 (33)
Unknown	2 (1)	2 (0)
Internet access		
Yes	248 (75.4)	
No	81 (24.6)	

Annual Fee

Of all patients with Internet access, 74.6% (n = 185) were willing to pay a small annual fee for one or more of the following online services: emailing with their physician, medication refills, viewing parts of their medical record, appointment requests, and billing inquiries. On a per service basis, 67% of patients with Internet access were willing to pay a "small fee" for "secure email" with their physician, 62% for online refills, 60% to review their medical record, 57% to request appointments, and 52% to make billing inquiries (Table 2). The differences in these responses were significant (P = .04; Fisher exact test, yes vs no vs no response 5x3 contingency table).

Table 2. Willingness of patients with Internet access to pay a small fee for specific online services (N = 248)

Service	Yes n (%)	No n (%)	No Response n (%)
Email with doctor	165 (67)	76 (31)	7 (3)
Medication refills	153 (62)	79 (32)	16 (6)
Viewing record	148 (60)	89 (36)	11 (4)
Appointment request	141 (57)	93 (37)	14 (6)
Billing inquiry	128 (52)	101 (41)	19 (8)

The majority of patients with Internet access in all age ranges were willing to pay a small fee for at least one of the five online or Web portal services (Table 3). Although this willingness to pay ranged from 60% for those in their 50s to 90% for those in their 30s, these differences were not statistically significant (P = 0.06; Fisher exact test, yes vs no response 8x2 contingency table).

Table 3. Willingness of patients with Internet access willingness to pay for at least one of the five services, by age (N = 248)

Age (years)	n	Yes	%
Less than 18	12	9	75
18–29	28	17	61
30–39	21	19	90
40–49	44	36	82
50–59	47	28	60
60–69	46	35	76
70 and older	40	33	83
Unknown	10	8	80
All	248	185	75

Of all respondents (N = 329), regardless of Internet access, 47.1% (n = 155) were willing to pay US $10 or more per year, with a median amount of US $20 (Table 4). Of those with Internet access (n = 248), 60.1% (n = 149) were willing to pay US $10 or more per year, and 31% (n = 46) were willing to pay US $50 or more per year. Of those who were disabled (n = 24), 29% were willing to pay US $10 or more per year.

Table 4. Maximum fee patients were willing to pay for online services (N = 329)

Amount (US $)	0	< 10	10	20	50	100	> 100
n	140	34	41	68	35	10	1

Online Services

As Table 5 shows, the three most important services to patients with Internet access (n = 248), in order, were emailing with their physician (34%), viewing their record online (22%), and medication refills (11%) ($P < .001$; Fisher exact test, most important vs not most important 7x2 contingency table). Note that 12% of those without Internet access (10/81) were still willing to pay for the service.

Table 5. Most important online service and willingness to pay

Most Important Service	All Patients (N = 329)	Patients With Internet (N = 248)
	n %	n %
Email	93 (28)	84 (34)
Record viewing	57 (17)	55 (22)
Medication refills	31 (9)	28 (11)
Multiple selections	21 (6)	18 (7)
Appointment requests	15 (5)	15 (6)
Billing information	0 (0)	0 (0)
No response	112 (34)	48 (19)
Willingness to pay for at least one service	**195 (59)**	**185 (75)**

Discussion

In the study practice, three quarters of adults are online and three quarters of those stated they are willing to pay a small annual fee for at least one of the five listed Web portal services. Even 12% of non-Internet users are willing to pay (presumably using the service through a friend or relative's Internet access or a public source like the library). Willingness to pay for Web portal services did not appear to vary significantly by age for those who already have Internet access. Nearly half of all patients and 60% of patients with Internet access were willing to pay at least US $10 per year for one or more of these services. Over 30% of patients with Internet access were willing to pay US $50 or more per year.

This study showed that no single online service stood out as the obvious favorite, but patient-physician email generated the highest level of interest, followed by online viewing of personal medical records, and online medication refills. No one chose billing inquiries as the service they valued the most. These findings are consistent with a February 2005 Harris Poll in which 80% of respondents indicated an interest in asking online questions of their physician, 69% in making online appointments, 69% in receiving test results online, and 67% in online medication renewal [20].

Like the 2002 Harris study that showed only 37% of online adults would be willing to pay for email with their physician [1], the February 2005 Harris Poll also reported that only 36% of online adults were willing to pay to send and receive emails from their doctor [20]. That stands in marked distinction to the finding here of 67% of patients being willing to pay for email with their physician. It is unknown whether the 2638 adults in the more recent Harris study had regular primary care physicians. It is conceivable that many did not and that people who have an established primary care physician relationship are more willing to pay for these services. Still another estimate of some patient willingness to pay for online patient-physician correspondence comes from an academic internal medicine practice in Colorado. In their study, 48% of all patients (both Web portal and non-portal users) were willing to pay for electronic correspondence with their physician. However, the amount they were willing to pay was small, at a mean of US $4 and median of US $2 per message [3].

A small annual fee could add up quickly if it could be collected. Ideally, it would be collected electronically at the time a user signs up and then annually thereafter. Just a US $10 per year fee for a practice of 2500 patients with 47.1% willingness to pay would translate into US $11775 per year of additional revenue for the involved physician. Web portals are relatively inexpensive and the current cost of one commercial product is about US $900 per physician per year [21]. A question this study did not address was whether patients are willing to pay both a small annual subscription fee and a per-transaction fee for e-messaging. This issue deserves further study.

A limitation of this study is that even though the top of the survey form stated “This is a completely confidential study,” and even though the author avoided discussing the survey with them, patients may have been more inclined to give answers they thought he wanted to hear because they were in the office for care that day. This theoretically could have biased the results toward more favorable responses regarding payment.

Although the author’s practice is felt to be fairly diverse and to adequately reflect the demographics of Tucson, no two practices are alike. If this practice is representative of Tucson, Arizona, can its findings be generalized to other communities? The author believes so, but it would be wise to establish this with similar studies conducted elsewhere. Evidence that the practice is representative lies in the Internet access statistic of 75.4%, which is consistent with a 2006 phone poll that reported 77% of US adults are now online [22]. And even if this study is representative of family medicine populations, would the findings for internal medicine practices differ significantly? Given the results of Table 3, which show 83% of patients over age 70 willing to pay for at least one of the five online services, a rate similar to other age groups, it appears likely that these results can be generalized to internal medicine as well.

A concern raised by this study is that one vulnerable, higher need population, the disabled, had relatively low Internet access (42%), and of those who had access, only 29% were willing to pay US $10 or more for online services. Since it is logical to think that people with medical disabilities would benefit more than most others by engaging in online services, the likely explanation for both the lower access and the lower “interest” is financial constraint – having to spend limited income on other needs. It would be useful to learn how interested disabled patients would be if these services were free for them.

What people say and what they do are not always the same. It would be most revealing to implement the services mentioned here with an annual fee of say, perhaps, US $15 and an email fee US $25 per message used to manage a clinical concern and see how many patients sign up for each.

Acknowledgments

This study was conducted with the author’s own resources.

Conflicts of Interest

None declared.

Appendix

Web Portal Survey

This is a completely confidential survey.

Patient’s age: ☐ under 18 ☐ 18–29 ☐ 30–39 ☐ 40–49 ☐ 50–59 ☐ 60–69 ☐ 70 and older

Patient’s sex: ☐ male ☐ female

Patient’s job: ☐ student ☐ employed ☐ unemployed ☐ disabled ☐ retired

Do you have Internet access? ☐ Yes ☐ No

Would you be willing to pay a small fee for the following services?

- Secure Internet viewing of your medical record ☐ Yes ☐ No
  (includes: Health summary, medication list, problem list, allergy list, lab results)
- Secure two way email with Dr. Adler ☐ Yes ☐ No

- Secure requests for medication refills (via Internet) ☐ Yes ☐ No
- Secure requests for appointments (via Internet) ☐ Yes ☐ No
- Secure email for billing questions / issues ☐ Yes ☐ No

If you answered yes to any of the above services, what is the MOST that you would be willing to pay per year?

☐ less than $10 ☐ $10 ☐ $20 ☐ $50 ☐ $100 ☐ more than $100

Please check the one service that would be MOST important to you:

(check only one)

☐ Secure Internet viewing of your medical record

☐ Secure two way email with Dr. Adler

☐ Secure requests for medication refills (via Internet)

☐ Secure requests for appointments (via Internet)

☐ Secure email for billing questions / issues

COMMENTS:

Multimedia Appendix

Presentation of Study Findings [PPT (MS PowerPoint) file, 72 KB - jmir_v8i4e26_app1.ppt]

References

1. Taylor H, Leitman R, editors. Patient/physician online communication: many patients want it, would pay for it, and it would influence their choice of doctors and health plans. Health Care News. 2002 Apr 10. (8) p. 1-3 URL: http://www.harrisinteractive.com/news/newsletters/healthnews/HI_HealthCareNews2002Vol2_Iss08.pdf#search=%22patient%20physician%20online%20communication%20many%20patients%20want%20it%2C%20would%20pay%22 [accessed 2006 Aug 25] [WebCite Cache ID 5IOYEhTsm]
2. ; Harris Interactive. Harris Interactive/ARiA Marketing Healthcare Satisfaction Study. Final Report. 2000 Oct. URL: http://www.harrisinteractive.com/news/downloads/HarrisAriaHCSatRpt.PDF#search=%22harris%20interactive%20Aria%20marketing%20healthcare%20satisfaction%20study%202000%22 [accessed 2006 Aug 25] [WebCite Cache ID 5IOYLVTUb]
3. Lin CT, Wittevrongel L, Moore L, Beaty BL, Ross SE. An Internet-based patient-provider communication system: randomized controlled trial. J Med Internet Res 2005 Aug 5;7(4):e47 [FREE Full text] [Medline: 16236699] [doi: 10.2196/jmir.7.4.e47]
4. Martin JC, et al. The Future of Family Medicine: a collaborative project of the family medicine community. Ann Fam Med 2004 Mar;2(Suppl 1) S3-32. [Medline: 15080220]
5. Barr M, Ginsburg J. The advanced medical home: a patient-centered, physician-guided model of health care. Policy monograph. American College of Physicians. URL: http://www.acponline.org/hpp/adv_med.pdf#search=%22advanced%20medical%20home%20a%20patient%20centered%20physician%20guided%22 [accessed 2006 Jan 22] [WebCite Cache ID 5IOZ1AQLE]
6. Babbott SF, Bigby J, Day SC, Dugdale DC, Fihn SD, Kapoor WN, et al. Redesigning the practice model for general internal medicine: a proposal for coordinated care. Policy Monograph. Society of General Internal Medicine. URL: http://www.sgim.org/BRPFinalReport71106.pdf#search=%22redesigning%20practice%20model%20for%20general%20internal%20medicine%20proposal%20for%20coordinated%20care%22 [accessed 2006 Jul 14] [WebCite Cache ID 5IOZVUzcv]
7. Katz SJ, Moyer CA. The emerging role of online communication between patients and their providers. J Gen Intern Med 2004 Sep;19(9):978-983 [FREE Full text] [Medline: 15333064] [doi: 10.1111/j.1525-1497.2004.30432.x]
8. Sands DZ. Help for physicians contemplating use of e-mail with patients. J Am Med Inform Assoc 2004 Jul;11(4):268-269 [FREE Full text] [Medline: 15252925]
9. Leong SL, Gingrich D, Lewis PR, Mauger DT, George JH. Enhancing doctor-patient communication using email: a pilot study. J Am Board Fam Pract 2005 May;18(3):180-188 [FREE Full text] [Medline: 15879565]
10. Liederman EM, Morefield CS. Web messaging: a new tool for patient-physician communication. J Am Med Inform Assoc 2003 May;10(3):260-270. [Medline: 12626378]
11. White CB, Moyer CA, Stern DT, Katz SJ. A content analysis of e-mail communication between patients and their providers: patients get the message. J Am Med Inform Assoc 2004 Jul;11(4):260-267. [Medline: 15064295]

12. Kittler AF, Carlson GL, Harris C, Lippincott M, Pizziferri L, Volk LA, et al. Primary care physician attitudes towards using a secure web-based portal designed to facilitate electronic communication with patients. Inform Prim Care 2004;12(3):129-138. [Medline: 15606985]
13. Martin Z. Online messaging opening clinical doors. Health Data Manag 2006 May;14(5) 26-8, 30, 32 passim. [Medline: 16722504]
14. Singer G. Some docs now make Web calls, Aetna, other insurers, starting to offer medical services online. South Florida Sun-Sentinel 2006 Apr 20;1A [FREE Full text] [WebCite Cache ID 5IOaDsM22]
15. Kowalczyk L. The doctor will e-you now: insurers to pay physicians to answer questions over the Web. The Boston Globe. 2004 May 24. URL: http://www.boston.com/business/articles/2004/05/24/the_doctor_will_e_you_now [accessed 2006 Aug 24] [WebCite Cache ID 5IOapLMsz]
16. MacDonald K. Online patient-provider communication tools: an overview. Oakland, Calif: HealthCare Foundation/First Consulting Group; 2003 Nov. URL: http://www.chcf.org/documents/ihealth/PatientProviderCommunicationTools.pdf [accessed 2006 Aug 24] [WebCite Cache ID 5IOaSJe2f]
17. Scherger JE. Communicating with your patients online. Fam Pract Manag 2004 Mar;11(3):93-94 [FREE Full text] [Medline: 23218299]
18. Endsley S, Kibbe DC, Linares A, Colorafi K. An introduction to personal health records. Fam Pract Mgmt 2006 May;13(5):57-62 [FREE Full text] [Medline: 16736906]
19. Jossi F. Customer service. Patients find a new age of conveniences in kiosks, portals and carts. Healthc Inform 2005 Aug;22(8):31-33. [Medline: 102499171]
20. ; Harris Interactive. Many nationwide believe in the potential benefits of electronic medical records and are interested in online communications with physicians. 2005 Mar 2. URL: http://www.harrisinteractive.com/news/allnewsbydate.asp?NewsID=895 [accessed 2006 Aug 24] [WebCite Cache ID 5IOXB0Krv]
21. Lowes R. Time to update your website? Med Econ 2006 May 5;83(9) TCP3-TCP4, TCP6-7. [Medline: 16749627]
22. ; Harris Interactive. Over three-quarters of all US adults–an estimated 172 million–now go online. The Harris Poll #41. 2006 May 24. URL: http://www.harrisinteractive.com/harris_poll/index.asp?PID=668 [accessed 2006 Aug 24] [WebCite Cache ID 5IObOdVrd]

Abbreviations

EHR: electronic health record
PHR: personal health record

submitted 25.08.06; peer-reviewed by S Ross, J Scherger, S Endsley; comments to author 11.09.06; revised version received 30.09.06; accepted 10.10.06; published 26.10.06.

Please cite as:
Adler KG
Web Portals in Primary Care: An Evaluation of Patient Readiness and Willingness to Pay for Online Services
J Med Internet Res 2006;8(4):e26
URL: http://www.jmir.org/2006/4/e26/
doi:10.2196/jmir.8.4.e26
PMID:

Original Paper

eHEALS: The eHealth Literacy Scale

Cameron D Norman[1], PhD; Harvey A Skinner[2], PhD, CPsych

[1]Department of Public Health Sciences, University of Toronto, Toronto, ON, Canada and Peter A. Silverman Global eHealth Program

[2]Faculty of Health, York University, Toronto, ON, Canada

Corresponding Author:
Cameron D Norman, PhD
Department of Public Health Sciences
University of Toronto
155 College Street, Room 586
Toronto, ON M5T 3M7
Canada
Phone: +1 416 978 1242
Fax: +1 416 978 2087
Email: cameron.norman@utoronto.ca

Abstract

Background: Electronic health resources are helpful only when people are able to use them, yet there remain few tools available to assess consumers' capacity for engaging in eHealth. Over 40% of US and Canadian adults have low basic literacy levels, suggesting that eHealth resources are likely to be inaccessible to large segments of the population. Using information technology for health requires eHealth literacy—the ability to read, use computers, search for information, understand health information, and put it into context. The eHealth Literacy Scale (eHEALS) was designed (1) to assess consumers' perceived skills at using information technology for health and (2) to aid in determining the fit between eHealth programs and consumers.

Objectives: The eHEALS is an 8-item measure of eHealth literacy developed to measure consumers' combined knowledge, comfort, and perceived skills at finding, evaluating, and applying electronic health information to health problems. The objective of the study was to psychometrically evaluate the properties of the eHEALS within a population context. A youth population was chosen as the focus for the initial development primarily because they have high levels of eHealth use and familiarity with information technology tools.

Methods: Data were collected at baseline, post-intervention, and 3- and 6-month follow-up using control group data as part of a single session, randomized intervention trial evaluating Web-based eHealth programs. Scale reliability was tested using item analysis for internal consistency (coefficient alpha) and test-retest reliability estimates. Principal components factor analysis was used to determine the theoretical fit of the measures with the data.

Results: A total of 664 participants (370 boys; 294 girls) aged 13 to 21 (mean = 14.95; SD = 1.24) completed the eHEALS at four time points over 6 months. Item analysis was performed on the 8-item scale at baseline, producing a tight fitting scale with $\alpha = .88$. Item-scale correlations ranged from $r = .51$ to .76. Test-retest reliability showed modest stability over time from baseline to 6-month follow-up ($r = .68$ to .40). Principal components analysis produced a single factor solution (56% of variance). Factor loadings ranged from .60 to .84 among the 8 items.

Conclusions: The eHEALS reliably and consistently captures the eHealth literacy concept in repeated administrations, showing promise as tool for assessing consumer comfort and skill in using information technology for health. Within a clinical environment, the eHEALS has the potential to serve as a means of identifying those who may or may not benefit from referrals to an eHealth intervention or resource. Further research needs to examine the applicability of the eHEALS to other populations and settings while exploring the relationship between eHealth literacy and health care outcomes.

(J Med Internet Res 2006;8(4):e27) doi:10.2196/jmir.8.4.e27

KEYWORDS

Internet; literacy; public health; psychometrics; quantitative evaluation

Introduction

How do we determine whether individuals have the capacity to engage with eHealth programs and interventions? Health practitioners, eHealth developers, and researchers alike need to know if electronic health tools are suitable methods for effectively promoting population health and aiding health care. An often unmentioned assumption underlying the deployment of eHealth resources intended for public consumption is that consumers have the skills to use such resources to their optimal level. Yet, with over 40% of US and Canadian adults having basic (or prose) literacy levels below what is considered necessary to optimally participate in civil society [1,2], it is unlikely that eHealth will provide population-level benefits as it requires much more than just prose literacy. Consumer-directed eHealth requires the ability to seek out, find, evaluate and appraise, integrate, and apply what is gained in electronic environments toward solving a health problem, or *eHealth literacy* [3]. This composite skill requires that people are able to work with technology, critically think about issues of media and science, and navigate through a vast array of information tools and sources to acquire the information necessary to make decisions.

Informed decision making requires that people can adequately access, understand, and process health information to meet their needs. Access refers both to the literal ability to access information resources like health websites, but also the quality of this access. This includes the quality of the technology (eg, Internet connection speed, hardware, software) and the conditions of use, such as whether people have the privacy or time to properly engage eHealth resources. Access in the Internet age also requires an ability to derive meaning from text. As basic literacy skills rise, so does the ability to use computers effectively to solve problems, regardless of age, income, or education [4].

Given issues of access and literacy, health practitioners in clinical and public health settings require an understanding of what abilities their patients/clients have before recommending eHealth resources. This article describes the development and psychometric evaluation of a measure of eHealth literacy designed for broad use in supporting consumer eHealth in public health and clinical care.

Health and Literacy in an Electronic Context

Health literacy has been identified as a public health goal for the 21st century and a significant challenge facing health care globally [5-7]. The recent Institute of Medicine report [8] on health literacy highlights the need to look at the different contexts where health information is obtained and used as part of a strategy of addressing health literacy. More than ever, this health information context includes electronic resources such as the World Wide Web and other technologies that now play an increasing role in consumer health [9,10]. Electronic health information introduces challenges pertaining to both the medium and the message that differ substantially from other media forms. Issues of access to information, retrieval, evaluation and appraisal, and other quality markers fundamentally differ in unregulated environments such as the Web, where new information is added every minute of every day. Being health literate in an electronic world requires a different or at least expanded set of skills to engage in health care and promotion, or eHealth literacy.

eHealth literacy is comprised of six core skills, or literacies: (1) traditional literacy, (2) health literacy, (3) information literacy, (4) scientific literacy, (5) media literacy, and (6) computer literacy [3]. The foundations of the eHealth literacy concept are based in part on social cognitive theory and self-efficacy theory [11], which promote competencies and confidence as precursors to behavior change and skill development and are described in detail elsewhere [3]. The challenge is developing the means to assess this skill in order to provide strategies to assist consumers in using eHealth to its fullest potential.

The eHealth Literacy Scale (eHEALS) has been developed to address the need to assess eHealth literacy for a wide range of populations and contexts. The eHEALS is a self-report tool that can be administered by a health professional and is based on an individual's perception of her or his own skills and knowledge within each measured domain. The instrument is designed to provide a general estimate of consumer eHealth-related skills that can be used to inform clinical decision making and health promotion planning with individuals or specific populations.

It is not unreasonable to assume a link between eHealth literacy and technology use in general. The more an individual uses technology, the more likely they are to develop skills in using that technology as a tool. For that reason, youth can serve as an ideal group to test a measure of eHealth literacy given this population's high familiarity with technology. In Canada, 99% of adolescents have access to the Internet, and the majority of Canadian teens report using the Internet for health in some capacity [12]. Although questions remain about the quality of this Internet access [13], this group is most likely to be familiar with information technology tools and is more likely to use eHealth than most other populations[13,14]. Despite having relative familiarity with eHealth, many adolescents are unable to derive the full benefit from it. Gray and colleagues looked at the issue of health literacy and technology in adolescents and found many teens experienced difficulty engaging with eHealth and understanding or using health information online, despite frequently using information technologies [15].

Regardless of the population of interest, the need to navigate the Internet with confidence is particularly important for health issues in which the consequences for using low quality, misleading, or false information are great [16]. By providing tools and resources to evaluate health information online and critically appraise eHealth resources, we offer an opportunity to both protect consumers from harm and empower them at the same time [17,18]. In order to provide relevant tools to aid consumers in navigating through eHealth, an understanding of what skills consumers possess at the outset, or their eHealth literacy, is required. This study's objective is to develop and test a functional method of assessing perceived eHealth literacy skills to aid consumers and health practitioners alike in assessing a fit with eHealth to support clinical care and promote population health.

Scale Development

A review of the literature was undertaken on each of the six key literacies that comprise Norman and Skinner's eHealth literacy model [3] in the Medline, PsycInfo, ERIC, Sociological Abstracts, and Web of Science databases to identify existing literacy measures. Although some measures were found, few had been rigorously psychometrically evaluated, and some were designed for specific projects that were not relevant to how the literacy concept was conceived of in relation to the eHealth literacy model. Given these constraints, it was decided that creating items from scratch was appropriate. Based on the theoretical model, an initial item pool was established and an iterative process of item reduction was used to create an instrument that could be easily deployed within a variety of settings and contexts as intended. The initial battery was circulated by the investigator to colleagues working in the area of eHealth for comment and review. After this initial review, the eHEALS was given to youth involved with TeenNet Research [19] to test general readability, item wording, and relevance. Youth are a consumer group with developing literacy skills and thus were expected to reflect the reading needs of a lower literacy population. These youth ranged in age from 12 to 19, came from many different social, ethnic, and educational backgrounds, and represented diverse interests among the adolescent population. Reviews were conducted in small groups over the course of 3 months. Further readability tests were conducted during the pilot phase of the project described below. Revisions were made as necessary before being pilot tested with a larger number of participants.

Pilot Testing

A total of 89 youth (ages 14-24) completed the initial, larger version of the eHEALS as part of a pilot test and provided comments on the readability and item wording in focus groups immediately following completion of the instrument in paper form. The eHEALS was subsequently reviewed and modified to create the final battery of 8 items based on the qualitative and response feedback from participants, theoretical fit, and comprehensiveness. This study represents its first full psychometric assessment.

Methods

Participants

This study was conducted as part of another larger evaluation project looking at eHealth smoking prevention and behavior change using a randomized controlled trial. The study described here involved participants from one arm of this trial given that the other arm was intended to promote eHealth literacy, thus potentially confounding the results of the psychometric review. The study recruited 664 adolescents from 14 secondary schools in a large Canadian city. Students in grades 9, 10, and 11 were sampled from a variety of class types encompassing different subject areas (eg, physical education, computer science) and formats (eg, single sex and mixed sex classes). An attempt was made to involve a cross-section of schools in the study through active recruitment directly with school administrators and teachers. Schools were offered a modest stipend for their involvement, but no direct incentives were provided to individual students as the study was considered a part of classroom activities due to a fit with the curriculum. Ethical approval for the study was obtained from the ethical review boards or committees from the University of Toronto, Toronto Public Health (a project partner), and both of the participating school boards.

Sample Characteristics

Age of the participants ranged from 13 to 21 (mean = 14.95; SD = 1.24), which included recent immigrants who may have been older than typical students in a particular grade. Sex was unevenly distributed within the sample due in part to the involvement of many single-sex classes involved in the study (boys N = 370; girls N = 294). The sample reflected the ethno-cultural diversity of the community, with the most commonly identified ethno-cultural groups being of Eastern European (N = 107, 16% of sample), East Asian (N = 103, 16% of sample), and Central Asian origin (N = 78), while 16% (N = 106) of participants did not identify with a particular cultural group. Ethno-cultural identity was determined using categories modified from Statistics Canada [2], which categorize individuals based on sociocultural and geopolitical differences in addition to racial ones. Most participants, 39%, were in grade 9 (N = 260), 29% in grade 10 (N = 193), and 32% in grade 11 (N = 211).

Technology Use

Participants reported being regular users of various forms of information technologies: 71% of participants (N = 468) reported using email at least once a week, with 37% (N = 544) using it daily; 79% (N = 522) reported using the Web each week, with 35% (N = 232) using it daily; and 71% (N = 473) of participants were regular (weekly) users of text messaging, with 42% (N = 280) reporting using it daily. Most participants reported that their primary access point for the Internet was at home (81%, N = 537); school (42%, N = 276) and friends' homes (34%, N = 227) were identified as the most common secondary access points for the Internet.

Procedure

The eHEALS was administered within a larger battery of measures as part of a combined randomized trial evaluation of an eHealth literacy promotion intervention and a Web-based smoking cessation program [20]. For the purposes of evaluating the eHEALS, data from the smoking prevention and cessation arm of the study (ie, the control condition) were used in the reliability testing of the eHEALS instrument. The eHEALS was administered using a pencil and paper survey delivered with other health measures used as part of a larger study. Participants completed the eHEALS prior to the intervention being delivered, immediately after the intervention, and at 3- and 6-month follow-up. Pre-test and immediate post-intervention data were collected during a single 75-minute class period.

Data Analyses

Internal consistency reliability was assessed using SPSS version 11.5[21] using the SPSS RELIABILITY command. Reliability (item) analysis was used to examine differences between boys and girls. Factors were identified using the simple structure approach solution based on reported eigenvalues over 1.0 [22]

using principal components analysis with SPSS FACTOR. This approach relies on a priori hypotheses to guide the selection of models, supported by scree tests and interpretability of the factor based on item/scale correlations. The results were considered using Comrey and Lee's (1992) guidelines whereby factor loadings in excess of .71 (50% overlapping variance) were considered excellent, .63 (40% overlapping variance) very good, and .55 (30% overlapping variance) good [23]. Factor loadings lower than .55 were considered fitting if items or scales correlated on only a single factor.

Results

Internal Consistency and Factor Analysis

The internal consistency reliability and factor analysis results are presented in Table 1. Each item in the eHEALS uses a 5-point Likert scale to answer each question with response options ranging from "strongly agree" to "strongly disagree" (Multimedia Appendix 1). Item analysis was performed on the 8-items, producing a tight fitting scale with coefficient alpha (α) of .88. Item-scale correlations between items ranged from r = .51 to .76. Principal components analysis was performed and produced a single factor solution as expected (eigenvalue = 4.479, 56% of the variance explained). Factor loadings ranged from .60 to .84 among the 8 items.

Table 1. eHEALS scale reliability and factor analysis

Item	Factor Loading	Mean Item-Total Correlation
Q1: I know how to find helpful health resources on the Internet	.77	.68
Q2: I know how to use the Internet to answer my health questions	.79	.70
Q3: I know what health resources are available on the Internet	.77	.68
Q4: I know where to find helpful health resources on the Internet	.84	.76
Q5: I know how to use the health information I find on the Internet to help me	.81	.73
Q6: I have the skills I need to evaluate the health resources I find on the Internet	.72	.63
Q7: I can tell high quality from low quality health resources on the Internet	.65	.55
Q8: I feel confident in using information from the Internet to make health decisions	.60	.51
Variance accounted for = 56%		
Coefficient alpha = .88		

Test-Retest Reliability

eHealth literacy scale scores were calculated and test-retest reliability was assessed by Pearson product moment correlation between scores at each interval (time 1 to time 4) using a standard regression model (SPSS REGRESSION) and using the intra-class correlation coefficient (SPSS RELIABILITY ICC MODEL (MIXED)). eHealth literacy scale scores were modestly correlated between administrations of the eHEALS, ranging from r = .49 to .68 (Table 2) . The intra-class correlation between the different scores was .49, suggesting that the eHEALS had modest stability over time.

Table 2. eHEALS test-retest reliability correlations

	eHealth Literacy Score Time 1	eHealth Literacy Score Time 2	eHealth Literacy Score Time 3	eHealth Literacy Score Time 4
eHealth Literacy Score Time 1	-			
eHealth Literacy Score Time 2	.68	-		
eHealth Literacy Score Time 3	.46	.49	-	
eHealth Literacy Score Time 4	.40	.40	.52	-

Relationship Between eHEALS and Other Measured Variables

Baseline levels of eHealth literacy were higher among males (t_{726} = 2.236, P = .026); however, no statistically significant differences were detected in scores at post-intervention and 3 and 6 month follow-up administrations of the eHEALS. Age did not predict eHealth literacy scores at any time point. No significant relationship was found between eHealth literacy and use of information technology overall or with respect to any individual forms of technology surveyed (WWW, TV, instant messaging, email, pager, or mobile phone) (P = .05). eHealth literacy levels were also not related to overall self-evaluations of health and were not a significant predictor of perceived health status over time in this sample.

Discussion

The eHEALS has shown promise as a measure of the concept of eHealth literacy as defined as a set of skills required to

effectively engage information technology for health. The eHEALS' high levels of internal consistency and modest test-retest reliability suggests that it has utility in examining eHealth literacy over time to both assess natural histories and evaluate eHealth literacy intervention outcomes. While tools exist that enable consumers to critically evaluate eHealth resources [24], there remains a dearth of instruments that assess consumers' skills at using eHealth in general. Indeed, relatively few validated measures exist for most of the key literacy conditions within the eHealth literacy model (eg, science literacy, information literacy). Thus, it is imperative that future studies examine the links between perceived skills, eHealth use, and health behavior and health outcomes. Of those literacy tools available, most require significant time resources to administer and analyze. The eHEALS was designed for simple, easy administration and thus can be used on its own or incorporated with other measures of health as part of a standard health assessment battery in primary care or to support health promotion planning.

eHealth literacy promotion takes place within a larger learning context, thus it makes sense to develop partnerships with other groups working within other literacy sectors in validating the eHEALS in relation to other measures of literacy, social functioning, health, and well-being. Two examples of such multi-sectoral partnerships include the National Literacy and Health Program sponsored by the Canadian Public Health Association [25] and the Learners Advisory Network of the Movement for Canadian Literacy [26], which brings literacy groups together to address systemic literacy issues. Such partnerships illuminate the shared challenges in creating capacity for research, development, and policy advocacy around health and literacy issues.

Limitations and Opportunities for Further Research

Conducting this study as part of a larger trial did pose problems for test-retest reliability; therefore, these results should be interpreted with caution. The lower than expected test-retest correlations between administrations of the eHEALS is attributed to a rise in eHealth literacy scores from baseline to post-intervention follow-ups, attributed to the smoking prevention intervention used in collecting the data [20]. Although unanticipated, one potential explanation for this increase is that the control intervention was designed based on the principles advocated by the eHealth literacy intervention itself (eg, user-friendly and audience-specific language, easy to read and navigate), which could have influenced participants' eHealth literacy scores. This may explain the relatively modest correlations (.68 to .70) compared with what was expected.

Additional studies are required to longitudinally examine the eHEALS in study conditions that are not susceptible to influence of the characteristics of a specific intervention. Testing the eHEALS with a population that has high rates of information technology presents a limitation; however, it also provides an opportunity to understand the robustness of the measure within a specific population. Further research needs to consider the eHEALS' application to other populations as well as groups with highly variable levels of technology familiarity.

The eHEALS measures consumers' *perceived* skills and comfort with eHealth, not the skills directly. The eHealth literacy model includes six types of literacy, and thus each skill would require independent measurement, such as rigorous usability tests of standard computer equipment for computer literacy and reading aloud text passages to assess basic prose literacy. For health practitioners and consumers alike, such detailed assessment would be problematic in practice; however, it is worthwhile considering ways to conduct such measures in the future.

Conclusions

The need for skills in seeking, appraising, and applying lessons learned through use of eHealth resources is common across ages, genders, and cultural groups, and thus the potential applicability of the eHEALS as a standard assessment tool for gauging eHealth literacy in health care is high. Assessing consumers' comfort in using eHealth allows for the identification of skill gaps and can better assist those with low comfort levels in taking advantage of the potential benefits that eHealth can afford. Doing so may foster development of tools that can meet these needs and aid in creating appropriate strategies for bridging the digital divide in consumer health care quality. Only by increasing the understanding of the disparities between available eHealth tools and consumers' abilities to use them can the necessary steps towards eliminating them be taken.

Acknowledgments

The authors thank Oonagh Maley for her assistance in the planning and delivery of this study. This study was supported by research grants from the Canadian Institutes of Health Research, Ontario Ministry of Health, Health Canada, and the Ontario Tobacco Research Unit.

Conflicts of Interest

None declared.

Multimedia Appendix 1

eHealth Literacy Scale [DOC (MS Word) file, 51 KB - jmir_v8i4e27_app1.doc]

References

1. ; National Center for Education Statistics. National Assessment of Adult Literacy (NAAL). a first look at the literacy of America's adults in the 21st century. Washington, DC: Institute of Educational Sciences, U.S. Department of Education; 2005. URL: http://nces.ed.gov/NAAL/PDF/2006470.PDF [accessed 2006 Nov 11]
2. ; Statistics Canada. Building on our competencies: Canadian results of the International Adult Literacy and Skills Survey, 2003. Ottawa, ON: Minister of Industry; 2005. URL: http://www.statcan.ca/english/freepub/89-617-XIE/89-617-XIE2005001.pdf [accessed 2006 Nov 11]
3. Norman CD, Skinner HA. eHealth Literacy: Essential Skills for Consumer Health in a Networked World. J Med Internet Res 2006 16;8(2):e9 [FREE Full text] [Medline: 16867972] [doi: 10.2196/jmir.8.2.e9]
4. Veenhof B, Clermont Y, Sciadas G. Literacy and digital technologies: linkages and outcomes. Ottawa, ON: Statistics Canada; 2005. URL: http://www.statcan.ca/english/research/56F0004MIE/56F0004MIE2005012.pdf [accessed 2006 Nov 11]
5. Rootman I. Literacy and health in Canada: is it really a problem? Can J Public Health 2003;94(6):405-7, 412. [Medline: 23061744]
6. Nutbeam D. Health literacy as a public health goal: a challenge for contemporary health education and communication strategies into the 21st century. Health Promot Int 2000;15(3):259-267. [doi: 10.1093/heapro/15.3.259]
7. ; Ad Hoc Committee on Health Literacy for the Council on Scientific Affairs, American Medical Association. Health literacy: report of the Council on Scientific Affairs. JAMA 1999 Feb 10;281(6):552-557 [FREE Full text] [Medline: 99144710] [doi: 10.1001/jama.281.6.545]
8. ; Institute of Medicine. Health literacy: a prescription to end confusion. Washington, DC: The National Academies Press; 2004. URL: http://www.iom.edu/?id=19750 [accessed 2006 Nov 11]
9. Madden M, Fox S. Finding answers online in sickness and in health. Washington, DC: Pew Internet & American Life Project; 2006 May 02. URL: http://www.pewinternet.org/pdfs/PIP_Health_Decisions_2006.pdf [accessed 2006 Nov 11] [WebCite Cache ID 5KNcaUDuD]
10. Norman CD, Chirrey S, Skinner H. Consumer perspectives on e-Health. In: Skinner H. Promoting health through organizational change. San Francisco, CA: Benjamin Cummings; 2002:315-334.
11. Bandura A. Self-efficacy: the exercise of control. New York: W.H. Freeman; 1997.
12. ; Environics Research Group. Young Canadians in a wired world: what are youth doing online, and what do their parents need to know?. Toronto, ON: Environics Research Group; 2001. Report No. pn4737 URL: http://www.media-awareness.ca/english/research/ycww/ [accessed 2006 Nov 11] [WebCite Cache ID 5KJxU11Qq]
13. Skinner H, Biscope S, Poland B. Quality of internet access: barrier behind internet use statistics. Soc Sci Med 2003 Sep;57(5):875-880. [Medline: 22736020] [doi: 10.1016/S0277-9536(02)00455-0]
14. Skinner H, Biscope S, Poland B, Goldberg E. How adolescents use technology for health information: implications for health professionals from focus group studies. J Med Internet Res 2003 Dec 18;5(4):e32 [FREE Full text] [Medline: 14713660] [doi: 10.2196/jmir.5.4.e32]
15. Gray NJ, Klein JD, Noyce PR, Sesselberg TS, Cantrill JA. The Internet: a window on adolescent health literacy. J Adolesc Health 2005 Sep;37(3):243. [Medline: 16109345] [doi: 10.1016/j.jadohealth.2004.08.023]
16. Eysenbach G. Infodemiology: The epidemiology of (mis)information. Am J Med 2002 Dec 15;113(9):763-765. [Medline: 22405313] [doi: 10.1016/S0002-9343(02)01473-0]
17. Eysenbach G. Rating information on the internet can empower users to make informed decisions. BMJ 1999 Aug 7;319(7206):385-386 [FREE Full text] [Medline: 99365211]
18. Eysenbach G, Jadad AR. Consumer health informatics in the Internet age. In: Edwards A, Elwyn G, editors. Evidence-based patient choice: inevitable or impossible. Oxford, UK: Oxford University Press; 2001:289-307.
19. ; TeenNet Research Program. Home page. URL: http://www.teennet.ca/ [accessed 2006 July 1] [WebCite Cache ID 5KJwLT07R]
20. ; Smoking Zine Program. Home page. URL: http://www.smokingzine.org/ [accessed 2006 July 10] [WebCite Cache ID 5KJwIcN9t]
21. ; SPSS Base [computer software]. Version 11.5. Chicago, IL: SPSS Inc; 2003.
22. Thurstone LL. Multiple factor analysis. Chicago, IL: University of Chicago Press; 1949.
23. Comrey AL, Lee HB. A first course in factor analysis. 2nd edition. Hillsdale, NJ: Lawrence Erlbaum Associates; 1992.
24. Discern Home page. URL: http://www.discern.org.uk/ [accessed 2006 Nov 11] [WebCite Cache ID 5GcfanyQ3]
25. ; National Literacy and Health Program. Home page. URL: http://www.nlhp.cpha.ca/ [accessed 2006 Nov 11] [WebCite Cache ID 5KJw4kmPP]
26. ; Learners Advisory Network of the Movement for Canadian Literacy. Home page. URL: http://www.literacy.ca/lan/lan.htm [accessed 2006 Nov 11] [WebCite Cache ID 5HRxY9VS5]

Abbreviations

eHEALS: eHealth Literacy Scale

submitted 05.01.06; peer-reviewed by J Jaffery, I O'Boyle; comments to author 16.05.06; revised version received 11.08.06; accepted 07.11.06; published 14.11.06.

Please cite as:
Norman CD, Skinner HA
eHEALS: The eHealth Literacy Scale
J Med Internet Res 2006;8(4):e27
URL: http://www.jmir.org/2006/4/e27/
doi:10.2196/jmir.8.4.e27
PMID:

Original Paper

Evaluating Common De-Identification Heuristics for Personal Health Information

Khaled El Emam[1], BEng, PhD; Sam Jabbouri[2], BSc; Scott Sams[3], BPAPM; Youenn Drouet[4], MSc; Michael Power[5], LLB

[1]University of Ottawa and CHEO Research Institute, Ottawa, ON, Canada

[2]School of Computer Science, Carleton University, Ottawa, ON, Canada

[3]Department of Geography and Environment, London School of Economics and Political Science, London, UK

[4]Departement d'Informatique et de Statistique, Faculte de Sciences Economiques et de Gestion, Universite Lumiere Lyon 2, Lyon, France

[5]Gowling Lafleur Henderson LLP, Ottawa, ON, Canada

Corresponding Author:
Khaled El Emam, BEng, PhD
CHEO Research Institute
401 Smyth Road
Ottawa, ON K1H 8L1
Canada
Phone: +1 613 797 5412
Fax: +1 613 741 0397
Email: kelemam@uottawa.ca

Abstract

Background: With the growing adoption of electronic medical records, there are increasing demands for the use of this electronic clinical data in observational research. A frequent ethics board requirement for such secondary use of personal health information in observational research is that the data be de-identified. De-identification heuristics are provided in the Health Insurance Portability and Accountability Act Privacy Rule, funding agency and professional association privacy guidelines, and common practice.

Objective: The aim of the study was to evaluate whether the re-identification risks due to record linkage are sufficiently low when following common de-identification heuristics and whether the risk is stable across sample sizes and data sets.

Methods: Two methods were followed to construct identification data sets. Re-identification attacks were simulated on these. For each data set we varied the sample size down to 30 individuals, and for each sample size evaluated the risk of re-identification for all combinations of quasi-identifiers. The combinations of quasi-identifiers that were low risk more than 50% of the time were considered stable.

Results: The identification data sets we were able to construct were the list of all physicians and the list of all lawyers registered in Ontario, using 1% sampling fractions. The quasi-identifiers of region, gender, and year of birth were found to be low risk more than 50% of the time across both data sets. The combination of gender and region was also found to be low risk more than 50% of the time. We were not able to create an identification data set for the whole population.

Conclusions: Existing Canadian federal and provincial privacy laws help explain why it is difficult to create an identification data set for the whole population. That such examples of high re-identification risk exist for mainstream professions makes a strong case for not disclosing the high-risk variables and their combinations identified here. For professional subpopulations with published membership lists, many variables often needed by researchers would have to be excluded or generalized to ensure consistently low re-identification risk. Data custodians and researchers need to consider other statistical disclosure techniques for protecting privacy.

(J Med Internet Res 2006;8(4):e28) doi:10.2196/jmir.8.4.e28

KEYWORDS

Privacy; confidentiality; HIPAA; security; data disclosure; ethics

Introduction

The adoption of electronic medical records (EMRs) is growing [1-5]. Researchers are increasingly turning to EMRs as a source of clinically relevant patient data. There are calls for EMRs to support secondary uses of this data for observational studies, such as epidemiologic and health services research [6]. On the other hand, a majority of patients, and the public in general, are concerned about unauthorized disclosure and use of their personal health information in an era of the EMR [7-11]. Furthermore, rates of medical identity theft have been increasing, and the risks are exacerbated with the use of EMRs [12].

Epidemiologic and health services research commonly proceeds without express consent from subjects. There are good reasons for this. It has been shown that requiring consent introduces biases in recruitment because those individuals who do not consent or who are difficult or impossible to request express consent from tend to be different on important characteristics than those who consent and are actually recruited. In some cases, the express consent requirements also increase the cost and duration of the research [13-25].

Excessive restrictions on researchers' access to identifiable health information is considered detrimental to society at large because many beneficial studies can not be done [26,27].

To safeguard privacy, often one of the requirements for waiving express consent is that the data be de-identified at the earliest opportunity [28]. This is important because there is evidence that individuals can be re-identified using common variables (such as zip code, date of birth, and gender) by linking to publicly available information [29,30]. In addition, identifiability is a key consideration for institutional research boards in deciding whether consent is required [31].

There are different methods for de-identification: statistical disclosure control (SDC) methods [32] and heuristic methods. In practice, SDC methods are not used that often [28,33]; therefore, we focus on heuristic methods. A heuristic approach to de-identification consists of rules about which variables to generalize (also known as aggregation) and which variables to exclude from a data set when it is disclosed. For example, under the US Health Insurance Portability and Accountability Act (HIPAA), two of the three de-identification methods stipulated in the Privacy Rule require the removal of potential identifying variables as defined in the Safe Harbor List and the Limited Data Set [34]. The Canadian Institutes for Health Research privacy guidelines provide examples of generalizing variables (eg, generalizing date of birth to age and generalizing geographic information) as a means to reduce identifiability [35]. Clinical researchers often follow heuristics to ensure that the data they collect and disclose are anonymized, for example, some assume that using initials and date of birth to identify subjects poses low risk of re-identification by those not involved in their study [28]. Various de-identification heuristics are used to decide which variables to exclude when pharmacy prescription records are released to commercial data aggregators [36].

De-identification by removing or generalizing variables from a data set necessarily results in loss of information and may hinder drawing accurate conclusions from that data [37]. The amount and criticality of that loss will depend on the specifics of the data set and the questions the data set is intended to answer. But most researchers would argue that variables, such as date of birth (or its generalization to age) and gender, are critical for many analyses, and geographic information (such as zip/postal codes) may also be necessary [38,39].

Given that there is potentially a high cost to using de-identification heuristics, it is essential to determine whether common de-identification heuristics used in practice today do indeed ensure that the risks of re-identification are low. If they do ensure low re-identification risk, then a case can be made for complying with these heuristics. If there is evidence that they do not ensure low re-identification risk, then the research community needs to consider alternative SDC methods as a means to de-identify data sets and reduce the need for excluding or generalizing important variables.

In this paper we evaluate whether common de-identification heuristics ensure a low level of re-identification risk across different data sets and sample sizes (since the risk of re-identification varies with sample size [32]). The common heuristics we evaluate are a union of a subset defined in the HIPAA Privacy Rule, currently practised in clinical research, and presented in privacy guidelines. If the heuristics ensure a consistently low level of risk, then one can have confidence in using them to de-identify any data set.

Categorization of Variables

It is useful to categorize variables in a research data set into the following set of mutually exclusive categories since each category is treated differently in the context of de-identification:

- **Identifying variables.** These are variables that can directly identify individuals, such as name, email address, telephone number, home address, social insurance number, and medical card number. Since these variables are obvious identifiers, if they are included, the data set is clearly not de-identified. In some cases, more than one identifying variable is needed to identify an individual uniquely. For example, the name "John Smith" appears 298 times in a search of the public telephone directory in Ontario. However, combined with a telephone number, the individual can be more easily identified uniquely.
- **Quasi-identifiers.** These are variables that do not directly identify an individual but can play an important role in indirect re-identification. One way in which quasi-identifiers can be used for re-identification is by linking to external databases containing identifying variables (record linkage). There are some quasi-identifiers that have been studied more extensively than others, such as gender, date of birth, and postal/zip code.
- **Nonidentifying variables.** Such variables may be, for instance, clinical and lab values. They are generally not useful for re-identification. For example, an indicator variable on whether an individual has pollen allergies would most likely be a nonidentifying variable.

It is common in disclosures of health data sets that the identifying variables are removed. We will therefore focus on risks from the quasi-identifiers.

Uniqueness and Re-identification

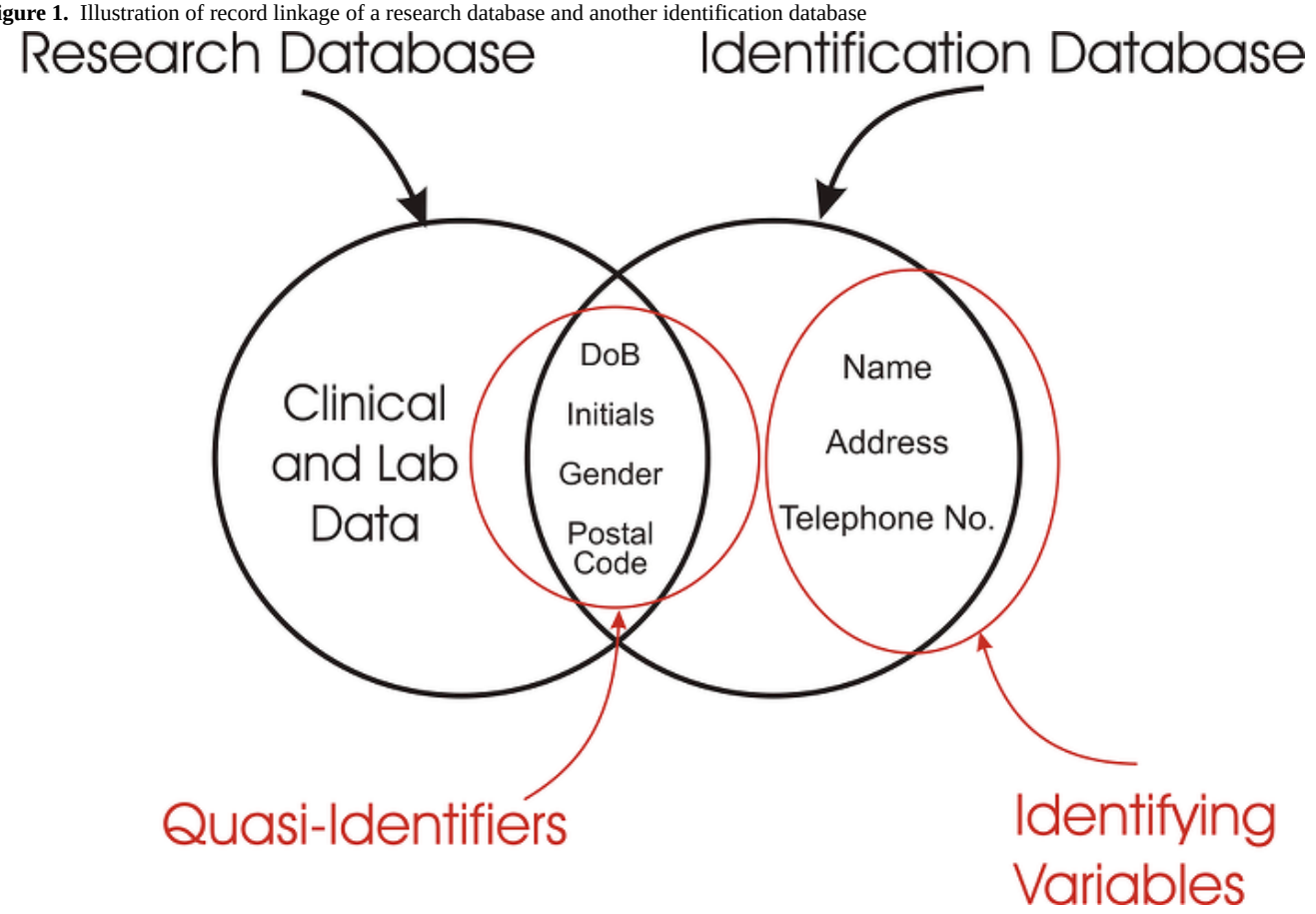
Uniqueness of individuals in a data set will have an impact on the risk of re-identification. We assume that a data set is a sample from some population. If an individual has a unique combination of values for the quasi-identifiers among other members of the population, then that person is *population unique*. If an individual has a unique combination of values for the quasi-identifiers among other individuals in the sample, then the individual is *sample unique*. If an individual is population unique, then, by definition, that person is sample unique, but not vice versa.

Uniqueness makes re-identification more likely through two common mechanisms: traceability and record linkage, which are explained below.

If a person is easy to trace in the real world, then that increases re-identification risk. For example, let's say that there are two quasi-identifiers in a data set: city/town and profession. If an individual has the values "Ottawa" and "Mayor," then it would be relatively easy to figure out who that individual is, even if there is no identifying information in the record.

If a particular set of quasi-identifiers in a record can be linked with a record in another database to re-identify individuals, then it can be said that the risk of re-identification is high. This is illustrated in Figure 1. Let us assume an individual (say, a researcher) has a de-identified research database containing some clinical data and that this database also contains quasi-identifiers such as initials, date of birth, gender, and postal code. If we could get access to an identification database or construct one from public data sources, with the same four variables as the research database as well as identifying variables such as name, address, and telephone number, then it would be possible to link the two databases and re-identify the individuals in the research database.

Figure 1. Illustration of record linkage of a research database and another identification database



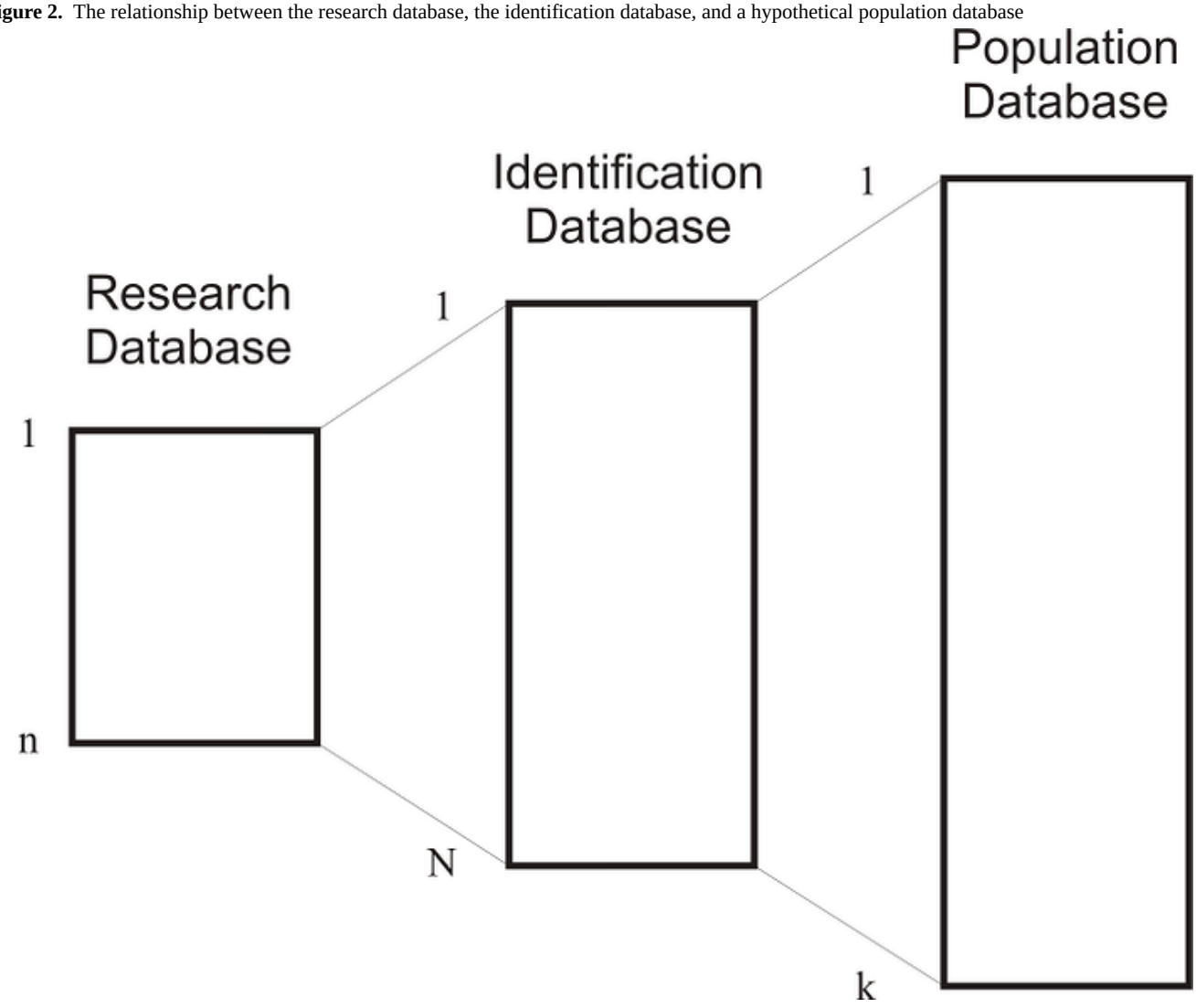
This means that if someone has access to a research database containing these quasi-identifiers, then it would be possible to re-identify the subjects by performing the record linkage with an appropriately constructed identification database. In principle, an identification database can be constructed in a number of ways:

- publicly available information from government bodies and professional associations
- data already available to an intruder from other sources, for example, a researcher with data available from another project (We will use the term "intruder" here for convenience, but it is recognized that re-identification may have legitimate purposes as well.)
- the circle of acquaintances of the intruder, which is the set of individuals from the population about which the intruder knows the values of the quasi-identifiers
- commercial organizations that sell databases containing data on members of the general public

- mining the Internet for information that individuals post about themselves (eg, resumes or personal Web pages) [40,41]
- inadvertent access to data, such as the purchase of surplus or second-hand computer equipment with data remaining [42]
- illegal activities, such as theft of computers with data or theft of unencrypted backup tapes during transit

Only the individuals in the identification database can be re-identified. If the identification database has all of the population of individuals in it (ie, $N = k$ in Figure 2), then all members of the population are potentially re-identifiable. The research database would represent a sample from the identification database, with $n < N$ in Figure 2. In the scenarios we are considering, an intruder is attempting to re-identify all the individuals in the research database.

Figure 2. The relationship between the research database, the identification database, and a hypothetical population database



Traceability and record linkage are two different things, although the underlying property (uniqueness) is the same, and one does not imply the other. For example, if we have a physician with a date of birth 1 January 1950 and that date is unique among all physicians in a province (ie, it is a population unique value), then that individual would still be difficult for an intruder to trace among the population of physicians. However, if an identification database of all physicians exists and the date of birth is one of the variables, then that physician would be easy to re-identify through record linkage.

We are only concerned with re-identification risk due to record linkage. Therefore, an important requirement is that an intruder is able to create an identification database. Only the quasi-identifiers that can exist in an identification database are relevant.

Commonly Used Quasi-identifiers

In the following paragraphs, we consider some of the commonly used quasi-identifiers in clinical research and their generalizations.

The first set of quasi-identifiers is defined in HIPAA. The HIPAA Privacy Rule defines three methods to de-identify a data set; two of these provide a very specific set of variables that should not be included in the data set for it to be considered de-identified. Both list a number of identifying variables and quasi-identifiers. We are only concerned with the quasi-identifiers. In the Safe Harbor method, two types of quasi-identifiers must be excluded:

- all geographic subdivisions smaller than a state (except the initial three digits of a zip code if the population in that zip code is more than 20000)
- all elements of dates (except year) or dates relating to an individual, including date of birth

The Limited Data Set method allows dates and excludes only the street address from the geographic information.

There is evidence that clinical researchers in Canada follow the HIPAA guidelines since these provide more precise prescriptions than anything else available locally [28]. Previous studies have performed successful matching experiments using date of birth, gender,and zip code as quasi-identifiers [29,30]. A recent qualitative study found that researchers use a combination of initials and date of birth to identify subjects [28]. Guidelines for protecting the privacy of personal information often include date of birth and geographic information as risky variables [35,43].

A generalization schedule for the geographic and date of birth information is as follows [35] (customized to a Canadian context):

- full postal code >> forward sortation area (first three digits of the postal code) >> city >> region (first character of the postal code)
- date of birth >> year and month of birth >> year of birth

A list of the quasi-identifiers extracted from the literature and evaluated in our study are given in Textbox 1.

Textbox 1. List of nine quasi-identifiers extracted from the literature

date of dirth (DoB)	forward sortation area
DoB – month and year	city
year of birth	region
gender	initials
postal code	

Methods

The objective was to evaluate re-identification risk for common quasi-identifiers and their combinations. The research method consisted of two steps:

1. constructing multiple identification databases
2. evaluating re-identification risk and its stability across data sets and sample sizes

Constructing Identification Databases

While there have been re-identification experiments in other nations, such as the United States [29,30], the United Kingdom [44], and Germany [45], there have been no attempts to construct identification databases in Canada. We therefore first attempted to construct identification databases using public sources in the province of Ontario.

Identifying Data Sources

Multiple sources of public data were sought as described below. Public data are defined as data that are available to the general public for free or a reasonable fee, with a reasonable amount of effort to access them, and without a review by the data holding institution or the need to sign a confidentiality or data sharing agreement with the data holding institution that restricts what can be done with the data.

All 29 Ontario government ministries were contacted. We identified staff in the freedom of information and privacy (FOIP) office in each ministry, if one existed. In all ministries except one, the FOIP office was contacted and we conducted a telephone interview with at least one staff member about the data that they release and the procedures for us to get that data.

A sample of commercial information brokers in Canada claiming to sell population databases were contacted to determine the type of data they hold, the sources of data, how the databases they sell were constructed from the sources, and conditions of disclosure. After examination of their websites, we followed up with phone calls to verify the information and get additional details. These brokers included Americanada, Prospects Influential, Nation Reach, and InfoCanada.

Sources of genealogical data were examined as well. These include data available through the Ottawa Public Library and the National Archive Centre. These include birth, baptism, death, marriage, adoption, and divorce data. Both of these locations were visited and staff on site were interviewed to determine the types of data available and how those data were released.

Professional societies frequently release comprehensive member lists. In some instances, work addresses and gender are also provided. We contacted a sample consisting of the College of Physicians and Surgeons of Ontario (CPSO), Law Society of Upper Canada (LSUC), Professional Engineers Ontario, College of Physiotherapists of Ontario, and the College of Occupational Therapists of Ontario. For all these professional societies, the membership lists were available on the Web. Commercial brokers, such as LexisNexis, WestLists, LawyerLocate, and Martindale, also provided lists of professionals. For commercial organizations, the data holdings were advertised on the websites. We followed up with phone calls to ensure the accuracy of the information on the Web and to fill in any missing details in our understanding of their data holdings.

We also contacted Statistics Canada and examined the information in the various products from the 2001 census data set. In particular, we focused on tabulations giving gender and age, and on microdata releases. Additionally, we contacted

Elections Canada and interviewed volunteers in election campaigns to understand how voter lists are used.

Creating Identification Databases

An identification database consists of two elements: quasi-identifiers and identifying information. There are two general methods that can be used for constructing an identification database:

- **Direct method.** A public source will have both elements needed for an identification database. An example would be a voters list.
- **Indirect method.** We first find a source with the identifying information on individuals, and then these are linked with another source that contains the quasi-identifiers.

We followed both methods to create an identification database.

Evaluating De-identification Heuristics

Measuring the Risk of Re-identification

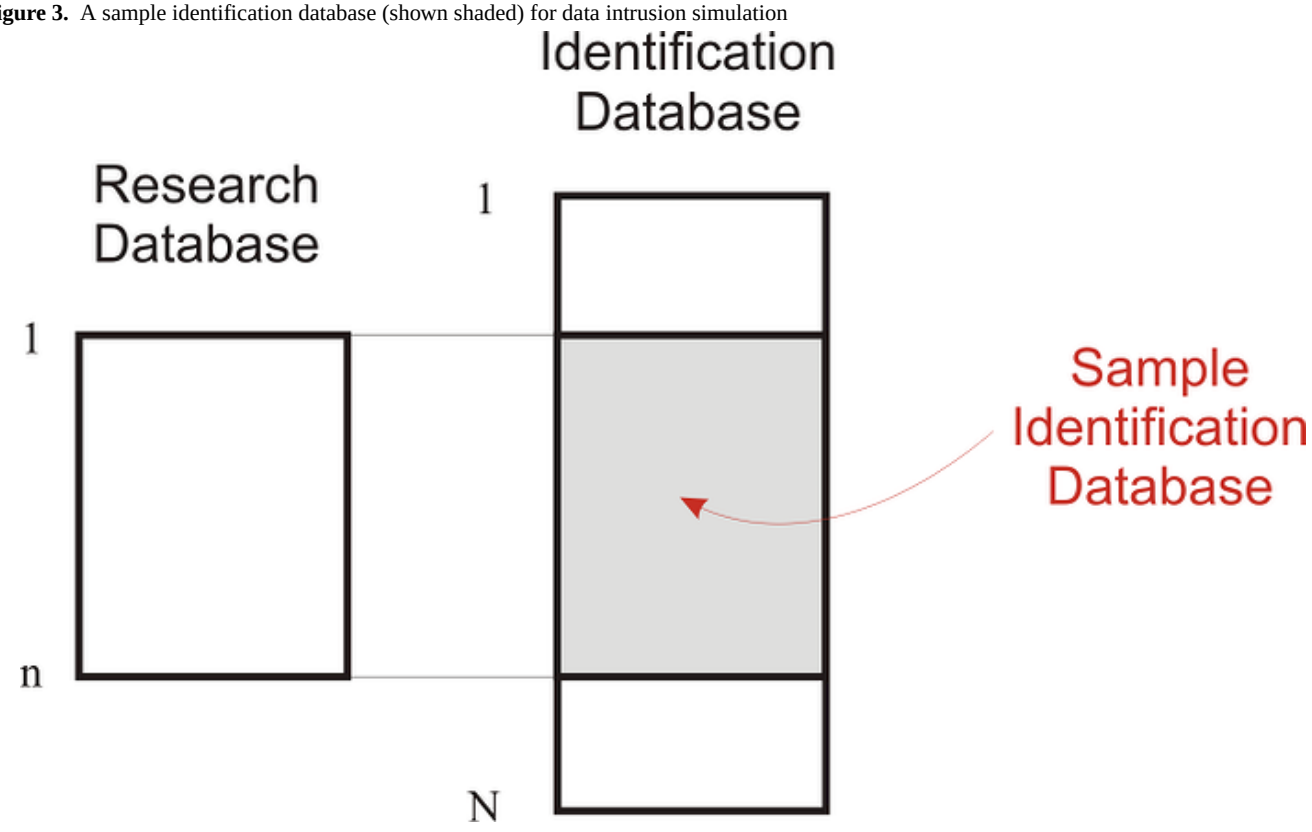
The measure of the risk of re-identification we used is grounded in the matching process that an intruder would likely use in order to re-identify a de-identified data set. Our measure of re-identification risk assumes that an intruder is attempting to re-identify all of the individuals in the research database by matching the individuals in the research database with records in an identification database using the quasi-identifiers. We predict the probability that a randomly selected individual can be matched successfully. Because only those individuals in the identification database can be re-identified, we assume that the identification database represents the population and the research database is a sample from that population (ie, only a subset of the individuals would be in the research database).

The estimation method used was data intrusion simulation (DIS) [46,47]. This predicts the risk of re-identification using this particular attack scenario (other attack scenarios are discussed later in the paper). DIS predicts the conditional probability that a unique match of a record in the identification database with a record in the research database is a correct match:

P(correct match|unique match) = P(cm|um).

It should be noted that we do not actually need a complete research database or a complete identification database to estimate re-identification risk. All that is needed is a sample identification database, as shown in Figure 3, containing only the quasi-identifiers and identifying variables for the *n* individuals in the research database. No actual clinical or lab data are required to perform the risk analysis.

Figure 3. A sample identification database (shown shaded) for data intrusion simulation



A Monte Carlo simulation, described in Appendix 1, illustrates the robust performance of DIS under a range of sampling fractions. Other measures of re-identification risk that have been proposed do not produce accurate results for small sampling fractions and are not specific to a type of attack [48,49].

Although there are no generally accepted re-identification thresholds, one can easily make the case that any probability of a successful attack greater than 0.01 would be unacceptable (for a large database, a probability of successful attack as high as 0.01 would compromise the privacy of a relatively large number of individuals). We therefore use that as a threshold for interpreting the risk results.

Evaluating the Heuristics

In our evaluation, three parameters were varied: the data set, the sample size, and the quasi-identifier combinations evaluated.

1. We constructed two identification databases to see whether the risk findings carried across them.
2. For each combination of quasi-identifiers, we decremented the sample size by one observation, chosen at random from *n* to 30, and determined whether *P(cm|um)* was below the threshold at the reduced sample size. This process was iterated 100 times for each sample size, and the average number of times that the risk was below the threshold was taken as the result for that sample size. If the risk was below the threshold, then we considered the quasi-identifier combination as "safe" (ie, one that ensures low re-identification risk quite often). We then looked at the frequency of quasi-identifiers that were considered "safe" across all sample sizes. If a quasi-identifier was "safe" more than 50% of the time, then it ensured that the risk was below the threshold across sample sizes.
3. We considered all possible individual and 2-, 3-, and 4-fold combinations of different quasi-identifiers.

Results

Constructing an Identification Database

Direct Method

The privacy offices at government ministries do provide oversight on the release of data. However, they are unable to control all possible releases and therefore only intervene when there is a complaint, an access to information request, or when they are asked for assistance from one of the departments. None of the privacy offices were able to produce a basic listing, even approximate, of all personal data releases from their ministry.

The commercial information brokers we contacted linked publicly available Statistics Canada census data with telephone directory data. Because of the aggregations performed on census data that are released, information such as age is only approximate. In addition, these would still not be population databases because not everyone has a telephone registered in their name. A recent independent study has confirmed that this is the approach used when commercial brokers utilize public data [25].

Birth and death notices are available from the General Registrar of Ontario. However, it is necessary to prove a relationship to the individual about whom data are being requested in order to get access to that information. Driver licence information also requires the name and the driver's licence number in advance to be able to make an information search request. Therefore, in both of these cases, it is not possible to construct a database for record linkage.

Voter lists are made available to candidates or their party representatives. These lists include the name, address, and date of birth of eligible voters. That information is to be used solely for the purposes of an election, including raising funds. Party members participating in an election campaign are bound by the party oath in terms of protecting that information. Volunteers on election campaigns who are not party members are not bound by an oath and would not normally sign a confidentiality agreement. Therefore, there are ways to get the voter list for a re-identification attack, but that would require deceptive practices and such use would likely go against the Elections Act.

Some commercial brokers may collect data sets directly from the public through surveys or subscription lists, or they may purchase these from retailers (eg, loyalty card users or warranty card information). These data sets may contain the quasi-identifiers we are interested in as well as identifying information. However, these do not include all members of the population.

We were therefore unable to construct an identification database for the whole population using the direct method.

Indirect Method

We were able to construct an identification database using the indirect method. However, it was not possible to do so for the whole population, but only for professional subpopulations, namely physicians and lawyers in Ontario. The list of physicians is published by the College of Physicians and Surgeons of Ontario (CPSO), and the list of lawyers is published by the Law Society of Upper Canada (LSUC).

It is possible to link the information in the list (which includes name, practice/firm address, and gender) with the Ministry of Government Services' Personal Property Security Registration (PPSR) data and the Canada 411 telephone directory data (both available on the Internet, the former for a fee) to identify the home postal code and date of birth (Figure 4).

We created a random sample data set of 236 physicians and 189 lawyers across Ontario with the quasi-identifiers under study. This represents a 1% sampling fraction of all registered physicians who are still active and practising in Ontario (23506) and all practising lawyers in Ontario (18728). The variables in our identification database were full name, gender, graduation date (CPSO only), date of birth, address for place of work (practice/firm), home address, and home telephone number.

Figure 4. The three main source databases used to construct an identification database for a professional subpopulation

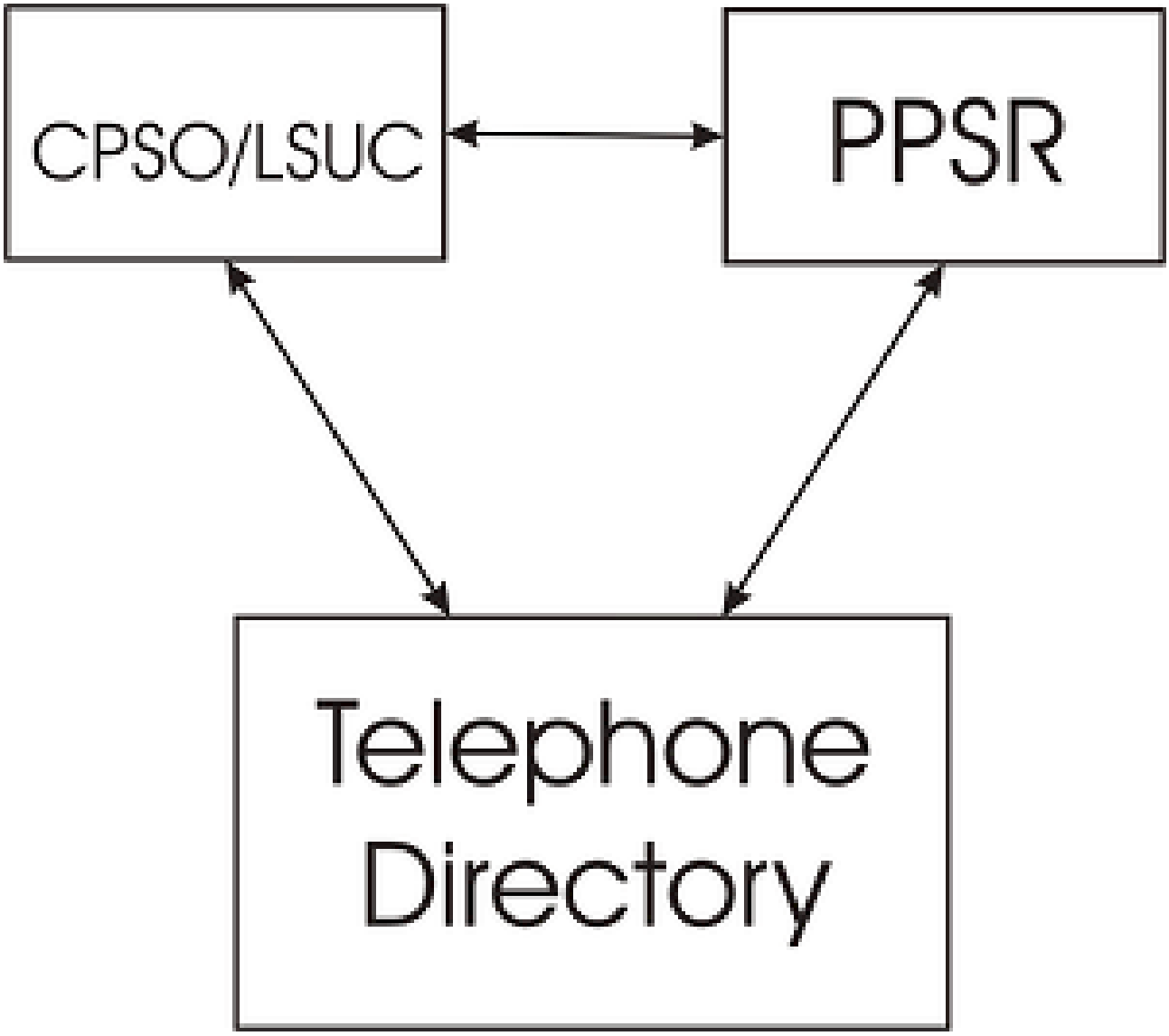


Table 1. Ability to get various data elements on physicians and lawyers, with the source of the data (n = 236 for CPSO; n = 189 for LSUC)

Quasi-identifier	CPSO (%)	LSUC (%)
home postal codes (source: PPSR and telephone directory)	60	45
practice/firm postal codes (source: CPSO/LSUC)	100	100
date of birth (source: PPSR)	40	45
gender (source: CPSO/genderizer for LSUC data)	100	100
initials (source: CPSO/LSUC)	100	100

Table 1 shows the success rates in getting the quasi-identifiers for an identification database. Name (and initials), practice postal codes, and gender are available from the CPSO. Therefore, we can obtain these for all physicians. Name and firm postal codes are available from LSUC. Since the LSUC does not publish gender in their public listing, genderizing software (see the analysis of the accuracy of such tools in Appendix 2) was used to estimate gender for the lawyers from their first names. We were able to determine the home postal code and date of birth from the PPSR for both professions. Additional verification of identity and home postal code was performed by checking against the Canada 411 website (online telephone directory). To verify that matches were correct, we also consulted the land registry in some instances to confirm addresses. Records were flagged for additional manual investigation under two conditions: (1) if the distance between the work and home postal codes was more than 100 km (determined by calculating the Euclidean distance), and (2) if, for physicians, the graduation date and date of birth were less than 25 years apart.

As evident in Table 1, it was not always possible to get the date of birth (40% and 45% success rates for physicians and lawyers,

respectively) and the home postal code (60% and 45% success rates for physicians and lawyers, respectively). There was also a gender difference. We were able to get the home postal code for 49% of all female physicians vs 63% for males, the date of birth for 29% of all female physicians vs 45% for males, the home postal code for 40% of all female lawyers vs 48% for males, and the date of birth for 40% of all female lawyers vs 48% for males.

Stability of Heuristics Across Sample Sizes and Data Sets

Table 2 provides the results of the stability analysis. The table shows the percentage of times that a particular combination of quasi-identifiers was found to be "safe" (ie, below the 0.01 risk threshold) as we varied the sample sizes across the two data sets. In total, 143 quasi-identifier combinations were evaluated. We only show those quasi-identifiers and their combinations that had percentages higher than 50%. If a quasi-identifier is not "safe" at least 50% of the time, then we can make the case that it is not stable. This means that if the quasi-identifier combination was above the risk threshold more than 50% of the time, it was therefore sensitive to sample size.

The findings indicate that gender, region, and year of birth are individually all relatively stable across sample sizes and data sets, as well as the combination of region and gender. This means that the inclusion of these quasi-identifiers in a released data set does not increase the risk of re-identification.

The gender and year of birth combination was low risk 80% of the time only for the CPSO data set. Consequently, we consider it unstable across data sets.

Table 2. Percentage of time a quasi-identifier or combination of quasi-identifiers was considered "safe" more than 50% of the time (as sample sizes were varied from 30 to the maximum)

Safe Quasi-identifier or Combination	Percentage of Time Quasi-Identifiers Were Below the Threshold	
	CPSO (%)	LSUC (%)
gender	100	100
region	93	65
DOB – year	94	85
gender + region	85	82
gender + DOB – year	80	–

Discussion

The Stable De-identification Heuristics

We found that only a small subset of the quasi-identifiers represented a consistently low risk of re-identification across both sample size changes and data set changes. Most quasi-identifiers (including generalizations) were not stable. In terms of formulating heuristics for the de-identification of data, the following quasi-identifiers were low risk (out of the set that we evaluated):

- region alone
- gender alone
- year of birth alone
- combination of gender and region

A corollary of this result is that all other individual quasi-identifiers and all other combinations are not safe.

Constructing Identification Databases

An important prerequisite for a record linkage attack is the ability to construct an identification database. It was possible to do so for professionals whose associations publish their membership lists. We found that it is more difficult to construct an identification database for adult females. It would also not be possible to perform a similar exercise on youth because youth would not have any loans that are registered, would not have property registered in their names, and would not have telephone numbers in their names. Therefore, their names would not appear in any of the publicly available data sources that we investigated. Also, it would not be possible to do so for professional associations that do not publish their membership lists.

We found that it is not possible to construct an identification database for the whole population of Ontario. We were unable to do so using public sources, with either the direct or indirect method. In Canada, the ability of researchers to access and use information is qualified by legislative restrictions designed to protect the privacy of individuals. This information may consist of what otherwise may be considered in other countries as "public data" (eg, driver's licence databases or public information).

In some instances, population databases are available for access but have certain data elements removed. For example, in Ontario, personal information is collected by the Ministry of Transportation under the authority of section 205 of the Highway Traffic Act. The information forms part of a public record and is used for the administration of the Ministry's driver, vehicle, and carrier programs. However, while residence address information is collected, it is not considered part of the public record and is not available to the general public. A further qualification is that only "authorized" requestors who have been approved and have entered into a contractual agreement with the Ministry may obtain residence address information for certain limited purposes. These purposes do include research by educational or research organizations. This limited degree of access is safeguarded by application of public sector privacy legislation in Ontario—the Freedom of Information and Protection of Privacy Act. The federal government and each of

the 13 provincial/territorial jurisdictions in Canada have similar legislation designed to protect the privacy of individuals and protect personal information held by government bodies.

Under such laws, "personal information" is broadly defined to generally mean recorded information about an identifiable individual, including "any identifying number, symbol or other particular assigned to the individual." Once it has been determined that a record contains personal information, these types of statutes generally prohibit the disclosure of this information, except in certain circumstances. One instance where disclosure may occur is when "personal information [is] collected and maintained specifically for the purpose of creating a record available to the general public," which is the case with the PPSR database we used.

The preceding discussion was directed to government holdings of information. The use of publicly available information held by non–public sector entities is governed by private sector privacy legislation that exists in Canada. At the federal level and in those jurisdictions that do not have comprehensive personal information protection statutes, the legislation in question is the Personal Information Protection and Electronic Documents Act. British Columbia, Alberta, and Quebec have their own statutes that place restrictions on the collection, use, and disclosure of personal information by non–public sector entities.

Generally, the provincial statutes governing non–public sector entities apply to publicly available information, making the use of such information subject to a consent requirement. Use without consent is permitted for certain prescribed sources of information. The federal statute permits the collection, use, and disclosure of publicly available information but then defines "publicly available information" by regulation. This include names, addresses, and telephone numbers in a telephone directory; name, title, address, and telephone number that appear in a professional or business directory available to the public; and personal information that appears in a registry collected under a statutory authority.

Generalization of Findings

Our data sets were constructed for an Ontario population. We have investigated the ability to construct similar identification databases in Canada. The two main data sources were the PPSR and telephone directory. There is an online telephone directory for every province. In Appendix 3, we have listed the PPSR sources for all provinces and territories. These would allow the construction of similar identification databases holding similar types of quasi-identifiers.

The risk of re-identification due to record linkage is affected by population uniqueness. For example, if we considered another profession that was heavily skewed toward males (say, underwater welders), then a female underwater welder is likely to be population unique. In that case, gender would not be a "safe" quasi-identifier. On the other hand, if there were no female underwater welders at all, then gender would be "safe." Notwithstanding such variations, our results provide concrete evidence that many common quasi-identifiers are high risk for some professions. That such examples exist for two mainstream professions makes a strong case already that the high-risk quasi-identifiers and combinations should not be disclosed.

As noted above, Canada has relatively strict privacy laws that restrict the amount of information about individuals that is disclosed and available for use in the public domain. Consequently, we expect that, from the perspective of re-identification risk, other jurisdictions with less restrictive laws would likely have higher risks of re-identification and more "unsafe" quasi-identifier combinations. Therefore, our list of "unsafe" quasi-identifiers is likely smaller than what one would find in a less restrictive jurisdiction (in terms of availability of information through public sources).

Given that the risk is affected by the ability to construct an identification database, this study can serve as a template for other jurisdictions to perform a risk assessment.

Managing the Risk of Re-identification

There are two ways to manage the risk of re-identification due to record linkage: exert control on the quasi-identifiers that are included in a research database, and exert control on the ability to create an identification database.

The first approach is simple to implement in practice. However, the quasi-identifiers that were found to be high risk constitute variables that would be considered important in many observational studies. These results highlight the unsatisfactory consequence of basing de-identification practices on such heuristics. This suggests that data custodians should consider using more sophisticated statistical disclosure control techniques [32] rather than basic heuristics about which variables to exclude and generalize. With such methods, it would be possible to retain important variables but at the same time reduce the risk of re-identification. This suggestion is essentially the third method defined in the HIPAA Privacy Rule for de-identifying data sets.

Two approaches to reduce the likelihood of being able to create an identification database are removing membership lists from the public domain and using financial deterrents.

Professional associations that make their membership lists public should re-evaluate this practice given the privacy consequences of doing so. The fact that such lists exist and are so easily accessible makes it possible to construct identification databases that can be used for launching re-identification attacks through record linkage. The most desirable action is to remove these lists from the public domain. Failing that, one would argue that at least the affected members should be made aware of the risks such disclosure entails.

When releasing membership lists it is also important to ensure that there are no unique values on all combinations of quasi-identifiers in the data set. The released data set would match the population, and population unique values represent a high risk of re-identification. For example, if we wish to release a list of all underwater welders and there is only one female, then that particular record should not be released, or the gender variable should not be released.

Another effective method is to impose fees for access to the registries that are used to create an identification database. Such

an access fee would be small enough to be, at most, an inconvenience to most legitimate users, but would represent a prohibitive cost for most intruders. There was a financial deterrent for the registries that were used in this study. At the time the study was conducted, there were 23506 physicians registered in Ontario who were still active and practising in the province. To be able to construct a complete identification database with records containing names, addresses (including postal codes), gender, and date of birth for all physicians in Ontario, it would cost at least Can $188048 because of the PPSR search fee (which is Can $8 per search). Similarly, there were 18728 registered lawyers, making the minimal cost for constructing an identification database Can $149824. While we needed only a 1% sample to estimate risk, an intruder would require a complete identification database for re-identification.

Limitations

In our study, we used a particular measure of the risk of re-identification. This measure assumes a particular attack scenario on the database. Our conclusions are limited to that attack scenario, but there are other possible scenarios of attack; for example:

- an intruder may already know that an individual exists in a research data set and wishes to identify the record belonging to that individual
- a specific individual or small number of individuals have unique characteristics in a released data set (eg, a specific diagnosis) and an intruder wants to identify these specific individuals in the data set

We did not consider these types of attacks, but they certainly would be important ones to investigate in the future. We also made the assumption that all individuals in the research database have the same probability of re-identification. Future work should consider re-identification risk at the record level. For instance, by knowing which specific records are high risk, they can be targeted for disclosure control actions. This would result in fewer distortions to a data set.

The threshold for high risk that we chose was arbitrary. There are no precedents for defining acceptable risk of re-identification for the release of personal health information; therefore, the risk threshold will have to evolve as our understanding of acceptable risk evolves. Furthermore, acceptable risk is not static. Society may get to accept higher risk in return for specific conveniences or personal benefits. Conversely, acceptable risk may decrease if there is a perception of abuse by custodians or there is a sharp rise in medical identity theft.

There may be a profession whose distribution of quasi-identifiers has many unique observations (eg, predominantly of a single sex or very sparsely distributed geographically). In such a case, the “safe” quasi-identifiers identified here may no longer be safe. Future research should investigate other public membership lists to determine the uniqueness of their quasi-identifier values to test the generalizability of these findings across professions.

Conclusions

One commonly used approach to protect data that may be disclosed for research purposes is to de-identify it. Specific heuristics for de-identification are included in, for example, the HIPAA Privacy Rule and various privacy guidelines. The heuristics stipulate that variables which present a high risk of re-identification (quasi-identifiers), for example, because they can be used in record-linkage attacks, should be removed or generalized. In this study, we examined such risks by evaluating the re-identification risk due to record linkage with common quasi-identifiers across different data sets and sample sizes.

It was not possible to construct an identification database for the whole population, but it was possible to do so for professional associations that publish their membership lists (eg, physicians and lawyers). Our results indicate that few quasi-identifiers are safe because they maintain the re-identification risk below a threshold. These are region, gender, and year of birth. The only combination that was consistently low risk was region and gender. That such examples of high re-identification risk exist for two mainstream professions already makes a strong case that many common quasi-identifiers should not be disclosed. However, one can also argue that many potentially important variables for researchers would be made unavailable.

These findings therefore indicate that the use of heuristics may be too restrictive and that data custodians should consider more sophisticated statistical disclosure techniques to ensure that important variables are retained in a data set while ensuring that privacy is maintained.

Acknowledgments

This research was partially funded by a joint grant from Ontario Centres of Excellence (CITO) and TrialStat Corporation under the Student Internship Program, as well as a grant from the Office of the Privacy Commissioner of Canada. We wish to thank the following people for reviewing earlier versions of this paper and providing extensive feedback: Sheila Chapman, Nour El-Kadri, Anita Fineberg, Debra Grant, and Mary Lysyk. We also wish to thank Latanya Sweeney for her suggestions and input during the initial design of this study. This program of research was approved by the Research Ethics Board of the Children’s Hospital of Eastern Ontario Research Institute.

Conflicts of Interest

Khaled El Emam is a co-founder of and has financial interests in TrialStat Corporation, which develops electronic data capture systems for clinical research.

Appendix 1

Monte Carlo DIS Simulation

In this appendix we report on a Monte Carlo simulation to demonstrate and evaluate the characteristics of data intrusion simulation. We used data on the 23506 physicians listed by the College of Physicians and Surgeons of Ontario as our population. We created random samples of various sizes from that population.

For the simulation we drew samples varying in size from 100 to 3000 individuals. A series of three quasi-identifiers were evaluated individually: gender, work postal code, and the work forward sortation area. Each sample size was drawn 1000 times and, in each case, the estimate of the probability of successful re-identification was calculated from the sample. Since we have the population data set as well, it was possible to compute the actual probability using the population data set. The bias of the predicted probability is computed by comparing it with the actual probability.

Figure 5. The probability of a successful re-identification attack (y-axis) for various sample sizes (x-axis) based on a Monte Carlo simulation

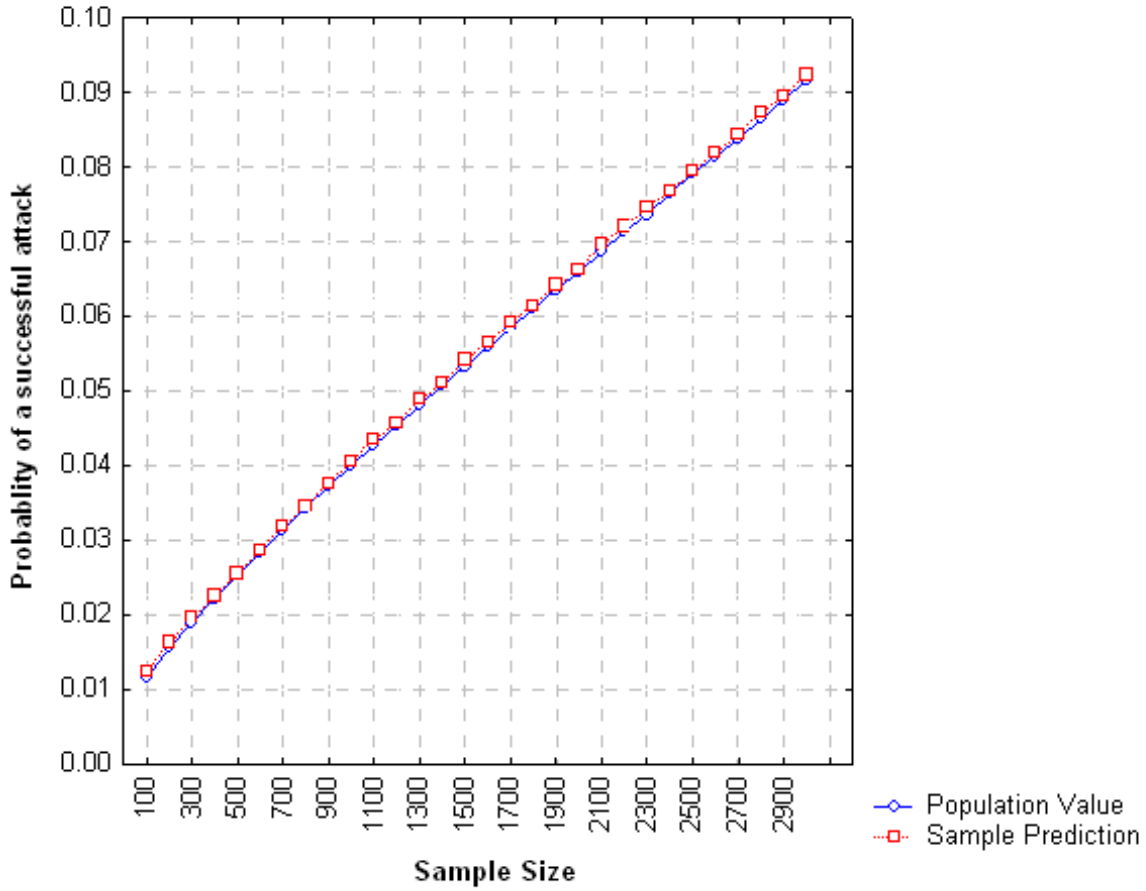


Figure 1 shows the results for the forward sortation area of the work postal code. The graph shows the predicted and actual values based on the sample and the population, respectively. These are averages across 1000 iterations for each sample size. It can be seen that the bias is quite small for the full range of sampling fractions studied. The highest sampling fraction was just under 13%, and the smallest sampling fraction was 0.04%. The magnitude of bias ranges from 0.00015 to 0.0016. For our purposes, this bias is quite small and indicates that the predicted probability is quite robust even for small sampling fractions. Similar results were obtained for the other quasi-identifiers examined.

It should also be noted that the probability of re-identification increases with sampling fraction. This is consistent with the common recommendation made in the disclosure control community to minimize the size of the sample that is released because that lowers the risk of re-identification. Therefore, one approach to reduce the risk of re-identification is to release a smaller data set.

Appendix 2

Evaluating Genderizing Software

The purpose of the evaluation described here is to find out how accurate current software is for predicting gender from first names. This capability is important in the construction of identification databases as a name is often available and we wish to obtain the gender quasi-identifier to match with a research database.

The data set that we tested with was the list of 23506 practising physicians in Ontario, for which we knew the correct gender. A search for genderizing software was performed on MedLine (no date limit), Journal of Marketing (2002 to present), Marketing (January 1996 to present), as well as Web searches on Yahoo and Google. The search terms used were “genderizer or genderizing or genderizing” and “software or tool or API”. Nine products were identified as well as the gender list provided by the US Census Bureau. Of the products, a number of them used the same underlying API. We contacted the vendors for the remaining products and were only able to successfully contact and purchase four products. The Census Bureau list is available for free. Each of the four products as well as the Census Bureau list was used to predict the actual gender for the list of physicians.

The results are shown in Table 1. The overall accuracy shown in the table is the simple proportion of overall predictions that were accurate. Precision, recall, and the f-measure are standard measures of binary classification accuracy. While the accuracy measures are quite high overall and tend to be quite close to each other, Personator with Genderbase had the best results for this data set. This is the tool that we used in our study to predict gender in the lawyer data set (LSUC). Given that this data set consists of heterogeneous Canadian professionals working in an Anglophone environment, it is reasonable to generalize to other similar groups. We cannot make broader generalizations to professionals, for example, in Francophone areas (eg, Quebec).

Table 3. Results of evaluating the accuracy of various tools for predicting gender from first names

	Male	Female
ParseRat (overall accuracy = 0.81)		
Precision	0.988	0.989
Recall	0.818	0.80
F-measure	0.89	0.88
Personator (overall accuracy = 0.81)		
Precision	0.98	0.99
Recall	0.82	0.79
F-measure	0.89	0.88
Personator with Genderbase (overall accuracy = 0.89)		
Precision	0.98	0.98
Recall	0.9	0.87
F-measure	0.94	0.93
MAILERS+4 (overall accuracy = 0.78)		
Precision	0.988	0.997
Recall	0.78	0.77
F-measure	0.87	0.87
US Census Bureau (overall accuracy = 0.77)		
Precision	0.98	0.996
Recall	0.77	0.78
F-measure	0.86	0.88

Appendix 3

Personal Property Security Registries

The following is the list of locations across Canada to obtain PPSR information for the construction of identification databases.

Table 5.

Province/Territory	URL
British Columbia	https://www.bconline.gov.bc.ca
Alberta	Available from authorized registry agents
Saskatchewan	http://www.isc.ca
Manitoba	https://direct.gov.mb.ca/ppr/
Ontario	https://www.personalproperty.gov.on.ca/ppsrweb/en/enquiry/cc_enquiry.jsp
Quebec	http://si2.rdprm.gouv.qc.ca/index.asp
New Brunswick	https://www.web11.snb.ca/snb7001/e/2000/2700e_6.asp
Nova Scotia	http://www.acol.ca/Services/PPR/NS/menu.html
Prince Edward Island	http://www.acol.ca/Services/PPR/PE/menu.html
Newfoundland and Labrador	http://www.acol.ca/Services/PPR/NF/menu.html
Northwest Territories	http://www.acol.ca/Services/PPR/NT/menu.html
Nunavut	http://www.acol.ca/Services/PPR/NU/menu.html

References

1. Irving R. 2002 Report on Information Technology in Canadian Hospitals. In: Canadian Healthcare Technology 2003.
2. ; Healthcare Information and Management Systems Society. HIMSS leadership survey. Trends in healthcare information technology. Healthcare Information and Management Systems Society Foundation. 2004. URL: http://www.himss.org/2004survey/docs/vendor_ceo_key_trends.pdf
3. Andrews JE, Pearce KA, Sydney C, Ireson C, Love M. Current state of information technology use in a US primary care practice-based research network. Inform Prim Care 2004;12(1):11-18. [Medline: 15140348]
4. Bower A. The diffusion and value of healthcare information technology. RAND Corporation. 2006. URL: http://www.rand.org/pubs/monographs/2006/RAND_MG272-1.pdf
5. Fonkych K, Taylor R. The state and pattern of health information technology adoption. RAND Corporation. 2005. URL: http://www.rand.org/pubs/monographs/2005/RAND_MG409.pdf
6. Powell J, Buchan I. Electronic health records should support clinical research. J Med Internet Res 2005;7(1):e4 [FREE Full text] [Medline: 15829476] [doi: 10.2196/jmir.7.1.e4]
7. ; Harris Interactive. Health information privacy (HIPAA) notices have improved public's confidence that their medical information is being handled properly. 2005. URL: http://www.harrisinteractive.com/news/allnewsbydate.asp?NewsID=894 [accessed 2006 Nov 12] [WebCite Cache ID 5KLoGrjNR]
8. ; Princeton Survey Research Associates. Medical privacy and confidentiality survey. Oakland, CA: California HealthCare Foundation; 1999. URL: http://www.chcf.org/documents/ihealth/topline.pdf
9. Grimes-Gruczka T, Gratzer C. The Institute for the Future. Ethics survey of consumer attitudes about health web sites. Oakland, CA: California HealthCare Foundation; 2000. URL: http://www.chcf.org/documents/consumer/Ethics2ndEdition.pdf
10. Willison DJ, Keshavjee K, Nair K, Goldsmith C, Holbrook AM, Computerization of Medical Practices for the Enhancement of Therapeutic Effectiveness investigators. Patients' consent preferences for research uses of information in electronic medical records: interview and survey data. BMJ 2003 Feb 15;326(7385):373 [FREE Full text] [Medline: 22473881] [doi: 10.1136/bmj.326.7385.373]
11. Mitchell E, Sullivan F. A descriptive feast but an evaluative famine: systematic review of published articles on primary care computing during 1980-97. BMJ 2001 Feb 3;322(7281):279-282 [FREE Full text] [Medline: 21096637] [doi: 10.1136/bmj.322.7281.279]
12. Dixon P. Medical identity theft: the information crime that can kill you. The World Privacy Forum. 2006. URL: http://www.worldprivacyforum.org/pdf/wpf_medicalidtheft2006.pdf [accessed 2006 Nov 20]
13. Melton III L. The threat to medical-records research. New England Journal of Medicine 1997;337(13):1466-1470. [doi: 10.1056/NEJM199711133372012]
14. Woolf SH, Rothemich SF, Johnson RE, Marsland DW. Selection bias from requiring patients to give consent to examine data for health services research. Arch Fam Med 2000;9(10):1111-1118 [FREE Full text] [Medline: 20566608] [doi: 10.1001/archfami.9.10.1111]
15. McKinney PA, Jones S, Parslow R, Davey N, Darowski M, Chaudhry B, PICANet Consent Study Group. A feasibility study of signed consent for the collection of patient identifiable information for a national paediatric clinical audit database. BMJ 2005 Apr 16;330(7496):877-879 [FREE Full text] [Medline: 15778230] [doi: 10.1136/bmj.38404.650208.AE]

16. Tu JV, Willison DJ, Silver FL, Fang J, Richards JA, Laupacis A, Investigators in the Registry of the Canadian Stroke Network. Impracticability of informed consent in the Registry of the Canadian Stroke Network. N Engl J Med 2004 Apr 1;350(14):1414-1421 [FREE Full text] [Medline: 15070791] [doi: 10.1056/NEJMsa031697]
17. Armstrong D, Kline-Rogers E, Jani SM, Goldman EB, Fang J, Mukherjee D, et al. Potential impact of the HIPAA privacy rule on data collection in a registry of patients with acute coronary syndrome. Arch Intern Med 2005 May 23;165(10):1125-1129. [Medline: 15911725] [doi: 10.1001/archinte.165.10.1125]
18. Baranek P, Black C, Fooks C, Maslove L, McGrail K. Data, data, everywhere…improving access to population health and health services research data in Canada. Centre for Health Services and Policy Research and Canadian Policy Research Networks. 2005 Apr. URL: http://www.cprn.com/documents/36948_en.pdf
19. Jacobsen SJ, Xia Z, Campion ME, Darby CH, Plevak MF, Seltman KD, et al. Potential effect of authorization bias on medical record research. Mayo Clin Proc 1999 Apr;74(4):330-338. [Medline: 99236649]
20. Nelson K, Garcia RE, Brown J, Mangione CM, Louis TA, Keeler E, et al. Do patient consent procedures affect participation rates in health services research? Med Care 2002 Apr;40(4):283-288. [Medline: 22016379] [doi: 10.1097/00005650-200204000-00004]
21. Al-Shahi R, Vousden C, Warlow C, Scottish Intracranial Vascular Malformation Study (SIVMS) Steering Committee. Bias from requiring explicit consent from all participants in observational research: prospective, population based study. BMJ 2005 Oct 22;331(7522):942 [FREE Full text] [Medline: 16223793] [doi: 10.1136/bmj.38624.397569.68]
22. Junghans C, Feder G, Hemingway H, Timmis A, Jones M. Recruiting patients to medical research: double blind randomised trial of "opt-in" versus "opt-out" strategies. BMJ 2005 Oct 22;331(7522):940 [FREE Full text] [Medline: 16157604] [doi: 10.1136/bmj.38583.625613.AE]
23. Ward HJT, Cousens SN, Smith-Bathgate B, Leitch M, Everington D, Will RG, et al. Obstacles to conducting epidemiological research in the UK general population. BMJ 2004 Jul 31;329(7460):277-279 [FREE Full text] [Medline: 15284154] [doi: 10.1136/bmj.329.7460.277]
24. McCarthy D, et al. What is the effect of state legislation requiring patient consent for use of medical records in research? In: Center for Health Care Policy and Evaluation 1999.
25. ; The Canadian Internet Policy and Public Interest Clinic. On the data trail: how detailed information about you gets into the hands of organizations with whom you have no relationship. A report on the Canadian data brokerage industry. Ottawa, ON: Canadian Internet Policy and Public Interest Clinic; 2006. URL: http://www.cippic.ca/en/news/documents/May1-06/DatabrokerReport.pdf [accessed 2006 Nov 18]
26. ; Academy of Medical Sciences. Personal data for public good using health information in medical research. 2006. p. 1-77 URL: http://acmedsci.ac.uk/images/project/Personal.pdf
27. Gordis L, Gold E. Privacy, confidentiality, and the use of medical records in research. Science 1980 Jan 11;207(4427):153-156. [Medline: 80081531] [doi: 10.1126/science.7350648]
28. El Emam K. Data anonymization practices in clinical research: a descriptive study. Ottawa, ON: Health Canada. Access to Information and Privacy Division; 2006. URL: http://www.ehealthinformation.ca/documents/HealthCanadaAnonymizationReport.pdf
29. Sweeney L. k-anonymity: a model for protecting privacy. International Journal on Uncertainty, Fuzziness and Knowledge-Based Systems 2002;10(5):557-570. [doi: 10.1142/S0218488502001648]
30. Ochoa S, Rasmussen J, Robson C, Salib M. Reidentification of individuals in Chicago's homicide database a technical and legal study.: Massachusetts Institute of Technology; 2001. URL: http://web.mit.edu/msalib/www/writings/classes/6.805/reid.pdf
31. Willison D. Academic REBs and governance of privacy, confidentiality and security in database research. Ottawa, Ontario, Canada: First Workshop on Electronic Health Information and Privacy Conference; Nov 2005.
32. Willenborg L, de Waal T. Elements of statistical disclosure control: Springer-Verlag; 2001.
33. Beach J. >Health care databases under HIPAA statistical approaches to de-identification of protected health information. DIMACS Working Group on Privacy/Confidentiality of Health Data. 2003. URL: http://dimacs.rutgers.edu/Workshops/Health/abstracts.html
34. Pabrai U. Getting Started with HIPAA: Premier Press; 2003.
35. ; Canadian Institutes of Health Research. CIHR best practices for protecting privacy in health research. Ottawa, ON: CIHR; 2005 Sep. URL: http://www.cihr-irsc.gc.ca/e/documents/pbp_sept2005_e.pdf
36. Zoutman DE, Ford BD, Bassili AR. The confidentiality of patient and physician information in pharmacy prescription records. CMAJ 2004 Mar 2;170(5):815-816 [FREE Full text] [Medline: 14993178] [doi: 10.1503/cmaj.1021826]
37. Fefferman NH, O'Neil EA, Naumova EN. Confidentiality and confidence: is data aggregation a means to achieve both? J Public Health Policy 2005 Dec;26(4):430-449. [Medline: 102709297] [doi: 10.1057/palgrave.jphp.3200029]
38. Bow CJ, Waters NM, Faris PD, Seidel JE, Galbraith PD, Knudtson ML, The APPROACH Investigators . Accuracy of city postal code coordinates as a proxy for location of residence. Int J Health Geogr 2004 Mar 18;3(1):5 [FREE Full text] [Medline: 15028120] [doi: 10.1186/1476-072X-3-5]
39. Bonner MR, Han D, Nie J, Rogerson P, Vena JE, Freudenheim JL. Positional accuracy of geocoded addresses in epidemiologic research. Epidemiology 2003 Jul;14(4):408-412. [Medline: 22726452] [doi: 10.1097/01.EDE.0000073121.63254.c5]

40. Sweeney L. AI technologies to defeat identity theft vulnerabilities. AAAI Spring Symposium on AI Technologies for Homeland Security 2005 [FREE Full text]
41. Sweeney L. Protecting job seekers from identity theft. IEEE Internet Computing 2006:74-78 [FREE Full text]
42. Garfinkel S, Shilat A. Remembrance of data passed: a study of disk sanitization practices. IEEE Security and Privacy 2003:17-27. [doi: 10.1109/MSECP.2003.1176992]
43. ; Interagency Confidentiality and Data Access Group. Checklist on Disclosure Potential of Proposed Data Releases. Washington, DC: Statistical Policy Office, Office of Information and Regulatory Affairs, and Office of Management and Budget; 1999. URL: http://www.fcsm.gov/committees/cdac/checklist_799.doc
44. Elliot M. Disclosure risk assessment. In: Doyle P, Lane J, Theeuwes J, Zayatz L, editors. Confidentiality, disclosure, and data access: theory and practical applications for statistical agencies. Elsevier; 2001.
45. Blien U, Wirth H, Muller M. Disclosure risk for microdata stemming from official statistics. Statistica Neerlandica 1992;46(1):69-82.
46. Elliot M. A new approach to the measurement of statistical disclosure risk. International Journal of Risk Management 2000;2(4):39-48.
47. Skinner G, Elliot M. A measure of disclosure risk for microdata. Journal of the Royal Statistical Society (Series B) 2002;64(Part 4):855-867. [doi: 10.1111/1467-9868.00365]
48. Chen G, Keller-McNulty S. Estimation of identification disclosure risk in microdata. Journal of Official Statistics 1998;14(1):79-95.
49. Zayatz L. Estimation of the percent of unique population elements on a microdata file using the sample. Bureau of the Census Statistical Research Division Report Series. SRD Research Report Number: RR-91/08. Washington, DC: US Bureau of the Census; 1991. URL: http://www.census.gov/srd/papers/pdf/rr91-08.pdf

Abbreviations

CPSO: College of Physicians and Surgeons of Ontario
DIS: data intrusion simulation
EMR: electronic medical record
FOIP: freedom of information and privacy
HIPAA: Health Insurance Portability and Accountability Act
LSUC: Law Society of Upper Canada
PPSR: personal property security registration

submitted 17.06.06; peer-reviewed by J Tu, N Fefferman; comments to author 04.07.06; revised version received 07.10.06; accepted 03.11.06; published 21.11.06.

Please cite as:
El Emam K, Jabbouri S, Sams S, Drouet Y, Power M
Evaluating Common De-Identification Heuristics for Personal Health Information
J Med Internet Res 2006;8(4):e28
URL: http://www.jmir.org/2006/4/e28/
doi:10.2196/jmir.8.4.e28
PMID:

Original Paper

To Track or Not to Track: User Reactions to Concepts in Longitudinal Health Monitoring

Jennifer S Beaudin[1], SM; Stephen S Intille[1], PhD; Margaret E Morris[2], PhD

[1]House_n, Massachusetts Institute of Technology, Cambridge, MA, USA

[2]Digital Health Group, Intel Corporation, Beaverton, OR, USA

Corresponding Author:
Jennifer S Beaudin, SM
House_n, Massachusetts Institute of Technology
One Cambridge Center, 4FL
Cambridge, MA 02142
USA
Phone: +1 617 452 5677
Fax: +1 617 225 0027
Email: jbeaudin@mit.edu

Abstract

Background: Advances in ubiquitous computing, smart homes, and sensor technologies enable novel, longitudinal health monitoring applications in the home. Many home monitoring technologies have been proposed to detect health crises, support aging-in-place, and improve medical care. Health professionals and potential end users in the lay public, however, sometimes question whether home health monitoring is justified given the cost and potential invasion of privacy.

Objective: The aim of the study was to elicit specific feedback from health professionals and laypeople about how they might use longitudinal health monitoring data for proactive health and well-being.

Methods: Interviews were conducted with 8 health professionals and 26 laypeople. Participants were asked to evaluate mock data visualization displays that could be generated by novel home monitoring systems. The mock displays were used to elicit reactions to longitudinal monitoring in the home setting as well as what behaviors, events, and physiological indicators people were interested in tracking.

Results: Based on the qualitative data provided by the interviews, lists of benefits of and concerns about health tracking from the perspectives of the practitioners and laypeople were compiled. Variables of particular interest to the interviewees, as well as their specific ideas for applications of collected data, were documented.

Conclusions: Based upon these interviews, we recommend that ubiquitous “monitoring” systems may be more readily adopted if they are developed as tools for personalized, longitudinal self-investigation that help end users learn about the conditions and variables that impact their social, cognitive, and physical health.

(J Med Internet Res 2006;8(4):e29) doi:10.2196/jmir.8.4.e29

KEYWORDS

User-computer interface; computers, handheld; ubiquitous computing; home monitoring; personal monitoring; personal tracking; personal health record; diaries; self-help devices; smart homes

Introduction

Background

Baby boomers, the cohort of adults born between 1946 and 1964, will contribute to a growing medical crisis in many industrialized countries. As demographics shift and lifespans increase, a larger percentage of adults will require medical care. The rising cost of medical procedures in combination with the greater numbers of people needing assistance will place an enormous strain on health care providers.

Ubiquitous computing and health technology researchers have responded to this developing medical crisis by proposing the use of home-based and wearable sensor technology to help people to assess their own health and that of their loved ones [1]. Research has already shown that health indicators typically monitored in clinical settings can be successfully deployed for longitudinal tracking in homes (eg, [2,3]). Numerous research efforts exist to develop systems that automatically detect activities of daily living (eg, [1,4-11]) and specific conditions, such as changes in gait. Sensors embedded in the home (and on

mobile devices [12,13]) are proposed to collect longitudinal and contextually sensitive data that can then be processed to automatically detect important changes in behavior patterns caused by the onset of illness. Such systems usually collect data continuously or when someone is engaged in a particular activity of interest, such as playing a computer game [14]. Commercial systems use a small number of sensors per dwelling, typically motion sensors, to monitor variation from baseline movement throughout a home in the hopes of detecting serious conditions that lead to immobility (eg, QuietCare from Living Independently Group). These systems are popular for monitoring elders living alone [15]. Although in some work authors have advocated the use of these novel technologies for personal health tracking [16-18], the focus of much of this prior work on the use of ubiquitous sensing in the home is on health monitoring geared toward the health professional.

Unfortunately, both clinicians and potential end users of monitoring systems can be skeptical of the systems technologists are proposing. The technology faces significant barriers to adoption (see [19] for a discussion). Clinicians, for example, typically do not request assessment until a problem arises, and some clinicians limit predictive testing because of time, expense, and fear their patients will overreact. End users have reservations such as fear of being diagnosed with an illness with no known cure, fear of tests, fear of stigmatization, and fears about privacy violation or behavior being judged by family or clinical caregivers.

Therefore, rather than developing systems that only monitor health status and provide data to clinicians, an approach that tightly integrates traditionally separate areas of monitoring, compensation, and prevention [19] may reduce fears and increase adoption of home health technologies among end users. Monitoring systems that provide data of interest directly to laypeople may be received more positively and be more rapidly adopted outside of laboratory settings than systems that only track metrics of interest to health providers.

Monitoring technologies that are more widely available, such as manual logging tools and diaries, have been found to be helpful in improving communication with patients [20] and reinforcing lifestyle counseling [21]. However, logging can be difficult for patients to maintain over time [22], and data originally deemed useful from a practitioner's perspective may not be easy for users to apply to everyday situations [23]. Tracked data may need to have immediate applications and be particularly relevant from a user's perspective for these systems to be kept in use long term.

Home monitoring technologies face another barrier to adoption, a classic “chicken and egg” evaluation problem. To make a (statistically) convincing argument that home monitoring systems can provide useful indicators of early onset of disease will require studies where the monitoring technology is installed in many homes for long evaluation periods, most likely months or years. To justify the cost of a sufficient number of installations, however, will require evidence of the preventive health value of the monitoring systems.

Prior work has shown that end users often strongly believe they should keep their own personal health records in addition to those kept by their medical personnel [24]. It has been proposed that such records would be enriched by including data and subjective reports collected outside of the clinic [25], but to date, no research has been done to suggest what end users would choose to track and what value, outside of the clinic, tracking would provide.

In this work, we therefore focus on ways of developing longitudinal home health monitoring systems that provide high perceived value to end users. Our primary question is whether “monitoring” systems can be designed that might be adopted by end consumers for personal use—even by people who would not characterize themselves as sick or in need of monitoring by health professionals or other caregivers. The information these systems collect might be used to develop novel forms of personal health records, which have recently attracted interest [25]. We present qualitative results from interviews with 8 health professionals and 26 laypeople who were asked to respond to mock data visualizations for novel longitudinal home health and activity tracking technologies.

The data display interviews described later in this paper were motivated by a set of exploratory interviews conducted for a project to develop a novel cognitive performance tracking technology. We interviewed 11 US-based professionals in aging and cognition, including a gerontologist, a nurse specializing in geriatrics, a home nurse, a cognitive psychologist, an occupational therapist specializing in geriatrics, three neurology researchers, and three neuropsychologists. The interviews were unstructured and lasted approximately 1-2 hours in length.

During these initial interviews, we observed that experts had not yet fully considered the breadth and depth of patient states and activities that emerging technology will be able to monitor outside of the clinical setting. Therefore, the experts, particularly the clinicians, had difficulty generating specific ideas on how their practices could take advantage of nonclinical monitoring of health status indicators other than the already familiar indicators such as heart rate, weight, and blood pressure. Nevertheless, some experts did propose new health status indicators that could be tracked outside of the clinic, and these suggestions motivated us to continue this area of inquiry. They included “catastrophic reactions to mundane activities” (eg, becoming unduly anxious when calculating a bill), ability to multitask during cooking, and changes in speed of interaction with appliances.

Objectives

Our exploratory interviews confirmed that we needed visual aids to help experts and laypeople understand how emerging technologies might be used for monitoring health-related status. Most clinicians and laypeople have had limited exposure to longitudinal tracking methods and devices. Clinicians, for example, are accustomed to evaluating patients based on periodic and limited clinic data. Laypeople may have experience using a bathroom scale, thermometer, or pedometer, but they are unfamiliar with the capabilities of emerging ubiquitous computing technologies. To address these barriers, we mocked-up data displays representing a variety of constructs for a hypothetical patient or family. These materials were used as a concrete focal point for structured interviews with medical

experts and laypersons. The data displays are not necessarily proposed concepts but are instead a mechanism used to elicit feedback about longitudinal tracking ideas. They are intentionally diverse and provocative and are used as probes to elicit detailed reactions and self-reflection during interviews. Our design criteria for the displays and strategies we used to achieve these goals are listed in Textbox 1.

Textbox 1. Interview goals and examples of strategies for participatory design stimuli used to meet those goals

Interview Goals	Examples of Strategies Used in Displays
Invite reflection on longitudinal monitoring of particular variables and outcomes.	Many displays consisted of sequences of steps and multiple timescales.
Invite focus on the output instead of the mechanism of monitoring/tracking.	Technology depictions or descriptions were not included; some displays explicitly suggested that data are collected manually.
Encourage participants to model how they would respond if they had tracking data, including how they would interpret outcomes and what follow-up investigations they would conduct, if any.	Displays represented accumulation of data as though the tracking tools had been in use for some time; displays put focus on action of reviewing data, instead of collecting data.
Encourage participants to think and talk about themselves and their personal concerns, values, and preferences.	Axes on graphs were often not labeled to avoid fixation on data values. Participants were encouraged to talk about what the display would look like for them.
Encourage discussion about underlying issues related to tracking, rather than restricting feedback to evaluation of a particular idea.	Multiple examples on one display could be quickly turned on and off; we deliberately restricted the set of metrics for each example to encourage brainstorming about additional metrics.

The displays were designed to encourage participants to role-play scenarios. Participants were instructed to envision themselves or their patients self-monitoring and analyzing the personalized health data. The displays were intended to illustrate capabilities of longitudinal monitoring (Textbox 2) to which participants often alluded but had difficulty discussing in more detail without stimuli.

Textbox 2. Home health tracking concepts that people may have difficulty discussing and relating to their personal situations and concerns

1. Data collected over time can reveal patterns of change.
2. Context can be used to interpret reasons for change.
3. Comparisons can be made with population norms, personal goals/estimates, and peers' values.
4. Quantitative data can be used in combination with qualitative data (eg, journal entries).
5. Multiple metrics can be applied to assess health and behavior change.
6. Data can be used to motivate by highlighting the extent of a problem or documenting progress.
7. Data can be used to problem solve and evaluate interventions.
8. Data can be subjectively reported or objectively observed.
9. Data can be reviewed at specific times and locations.
10. Data can be organized in ways other than by time.
11. Data tracking may not be constant, instead triggered by directed investigations.
12. Data can be reviewed in isolation or in relationship to other variables.

Methods

We developed 17 mock data display examples, many of which represented multiple tracking concepts. Initially, the displays were developed on paper, but to facilitate long-distance interviewing, they were converted to interactive Web pages using Flash (Multimedia Appendix 1). Most of the examples could be developed using emerging ubiquitous computing and/or wearable technology. A few (eg, an example that assumes the computer can detect linguistic “hedges”) would not be possible to implement at this time. Four of the displays are described in more detail below. We also developed sorting exercises to provide participants with a way to express tracking priorities, complementing the more open-ended feedback from the displays; they are described in the final Methods subsection, “Sorting Exercises with Laypeople.”

Weight Displays Example

The weight display series was used to address possible monitoring strategies to facilitate a weight loss goal. The complexity increases sequentially. In the first display, one data point is available: the patient's weight at the clinic. This value is compared with norms for gender and height. The second display (Figure 1a) adds the patient's weight from a clinical visit one year ago, indicating that the currently overweight patient

has actually succeeded in significant weight loss. The third display (Figure 1b) suggests that the patient has been recording his or her weight each month throughout the year; the patient's weight was steady, started to steeply decline, and finally reached a plateau. The seasonal context indicates that decline began in January. The next in the series of displays (Figure 1c) switches metrics to show pedometer readings plotted against exercise goals for each week. The final display (Figure 1d) depicts the addition of qualitative description: photos taken on “good days,” when the user's steps greatly exceeded the goal for the given week, and “bad days,” when the user's steps significantly failed to meet the goal for the given week. These “photos” are left blank, but the participants are told that they could capture anything experienced on those days: people encountered, food eaten, places visited.

Figure 1a. Weight displays: (a) traditional clinical weight data

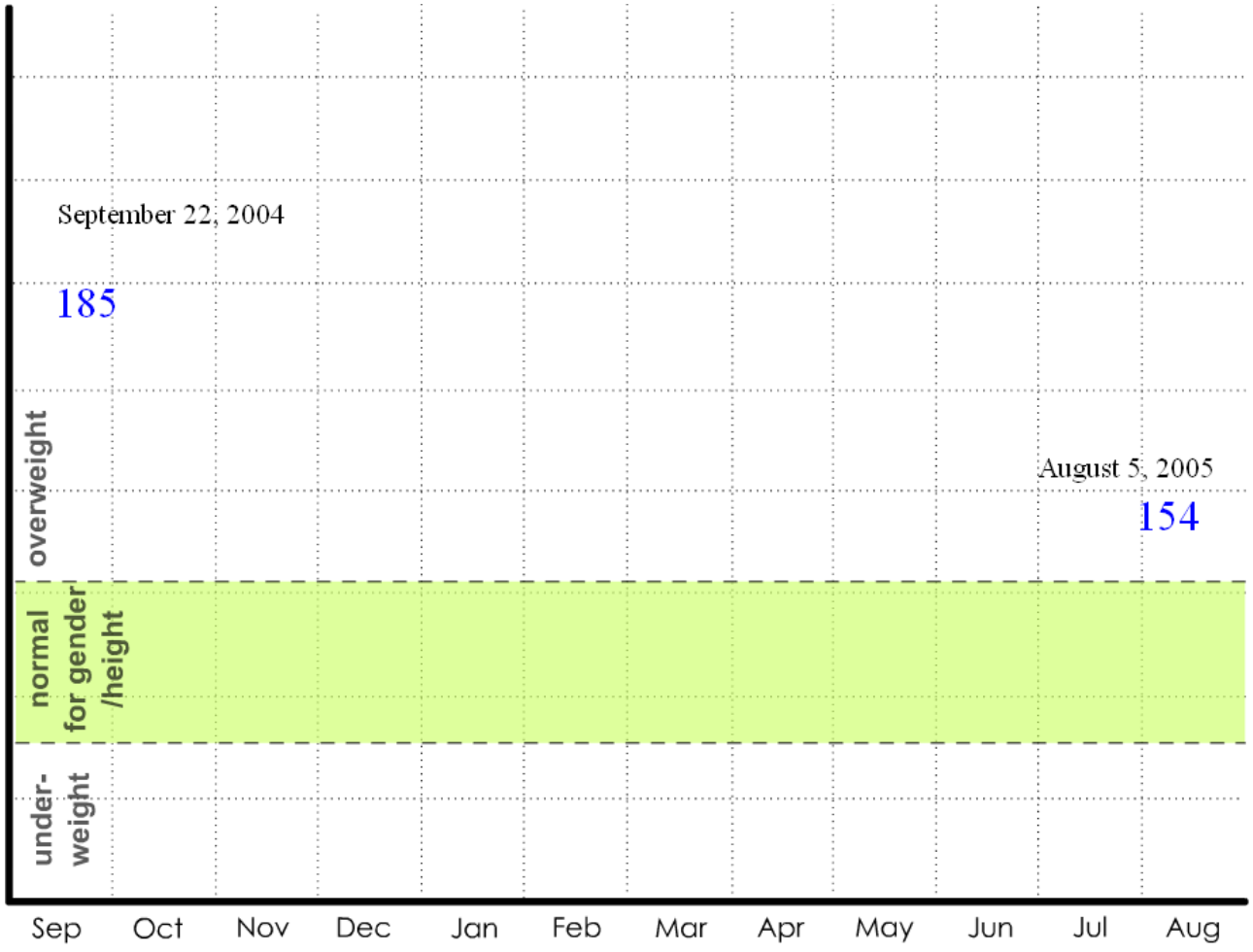


Figure 1b. (b) home monitoring weight data

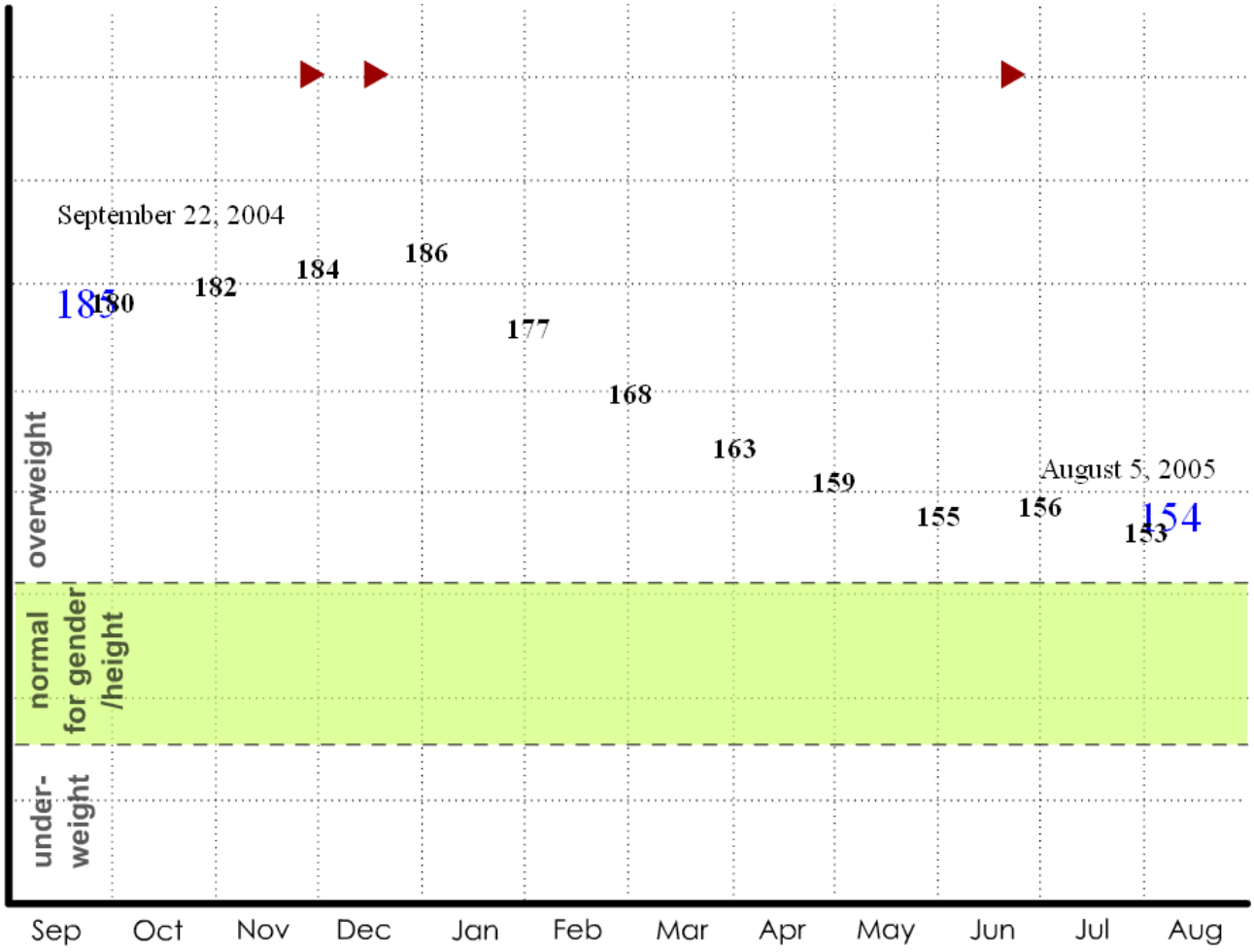


Figure 1c. (c) pedometer readings

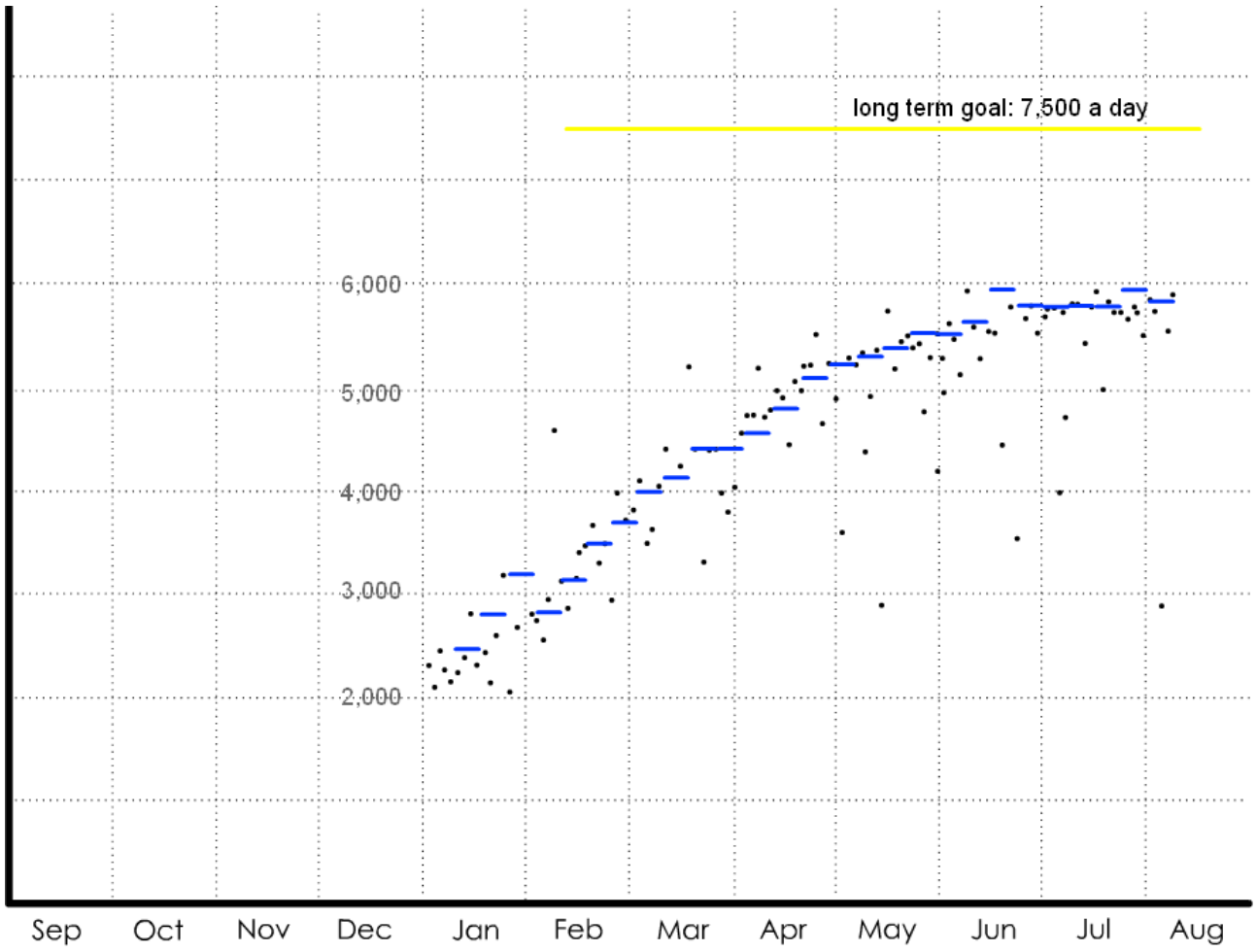


Figure 1d. (d) qualitative data—photographs

The displays tell a story that begins with the more familiar and accepted clinical metrics and gradually transitions to richer and more diverse sources of data, concluding with a qualitative display that could provide information about a variety of life factors. Displays such as these were used to help participants understand the potential of home health tracking technologies in a step-by-step fashion. In this example, the interviewer depicted a typical patient struggling with weight who committed to change (New Year's resolution), used a pedometer to record and motivate progress, and finally reached a plateau—a point which drives many back to poor habits because their efforts are not reinforced. The interviewer described how extra data, such as the analysis of contextual factors and qualitative image data, could help the patient-physician pair work through this apparent impasse. The tools for this weight example are all commonly available: a bathroom scale, a pedometer, and a digital camera. This display series illustrates concepts 1-7 in Textbox 2.

Scenario Displays Example

The weight displays example was developed specifically to help health professionals transition from familiar patient charts to more novel tracking concepts. To achieve a similar goal with laypeople, scenario displays (Figure 2) were developed showing extensions of familiar media associated with recording and reflecting on personal data. With the journal display (Figure 2a), participants were asked to imagine that they had the time and discipline to keep a richly detailed journal, which might include information as diverse as memorable moments and snacking events. With the health snoop display (Figure 2b), participants were asked to imagine that they could hire someone to follow them around for a week, nonintrusively making observations about such things as work habits and missed opportunities for exercise. These examples explicitly sidestep the data collection issues in favor of focusing on the basic ideas of subjective and objective recording and analysis of behavior (Textbox 2, #8). Other displays included data overlaid on a grocery receipt (Figure 2c), overlaid on a medicine cabinet mirror, and communicated by email or voicemail. These examples focused the participant on the action of reviewing and applying data (Textbox 2, #9), rather than on the collection of data.

Figure 2a. Scenario displays: (a) journal display

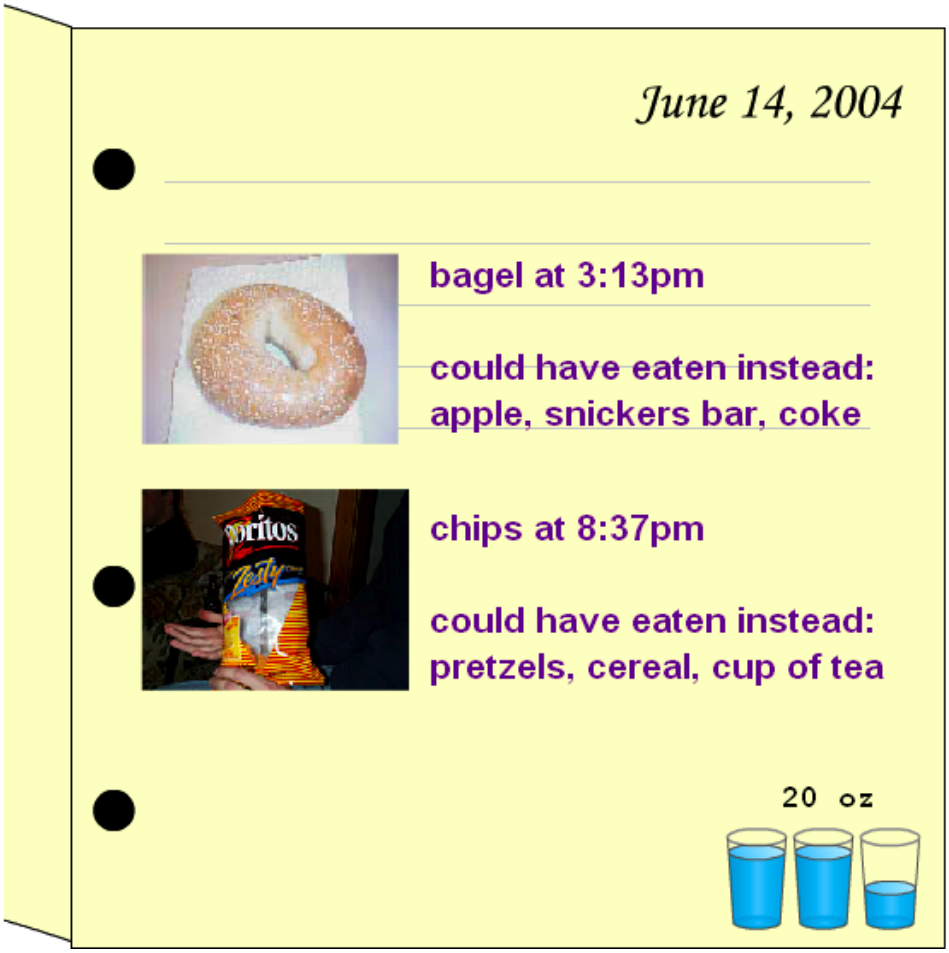


Figure 2b. (b) health snoop display

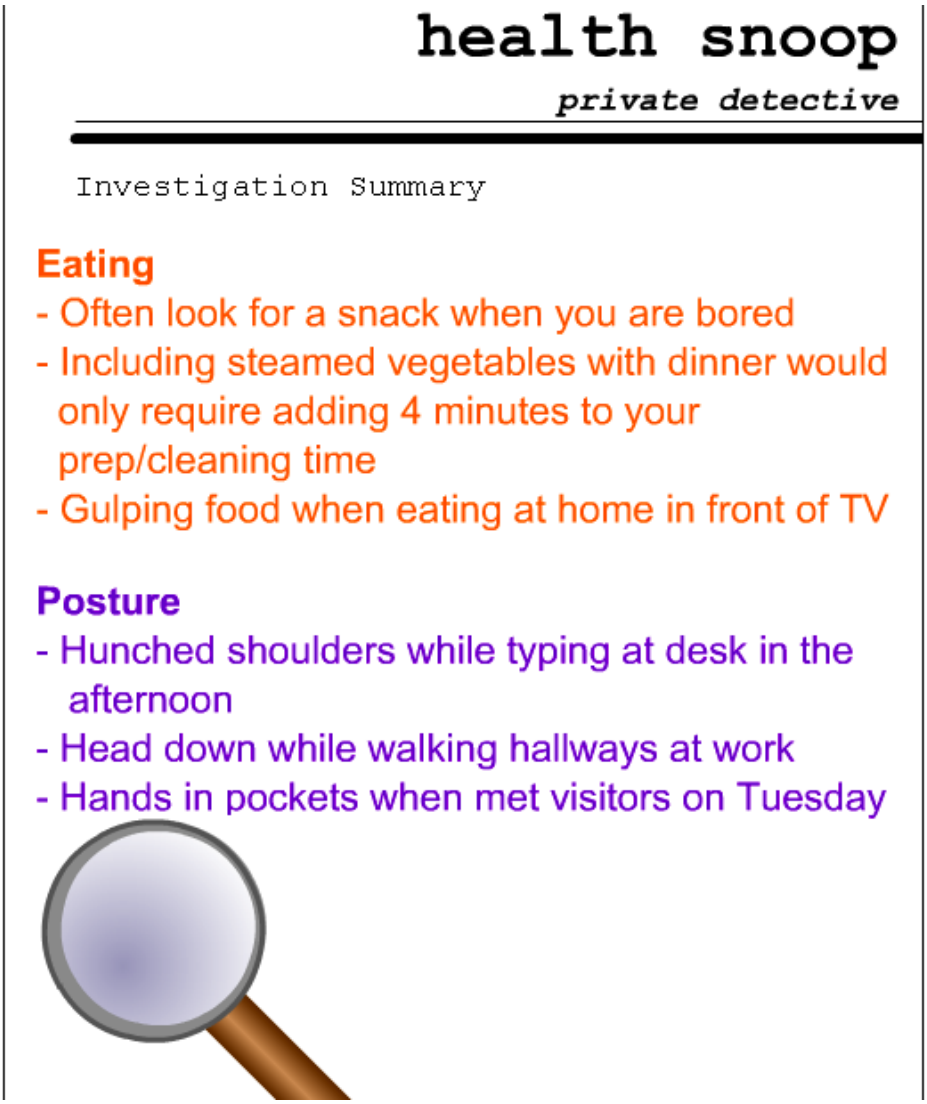
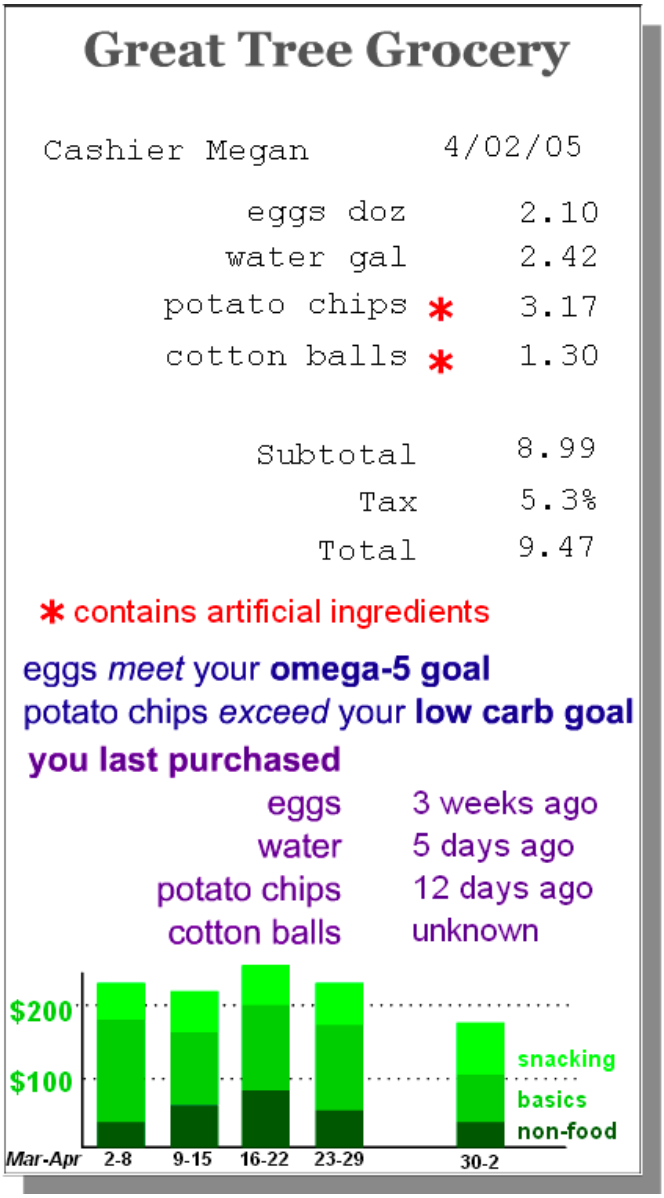


Figure 2c. (c) grocery receipt display



Body-Based Displays Example

The Web-based displays permitted customized interviewing, where a basic display template could be overlaid with multiple data examples by checking options on the screen. The body example (Figure 3) presented tracking data organized according to associated body regions (Textbox 2, #10). By checking the options in the bottom right corner and clicking on parts of the body, the participant could review ideas for tracking skin changes, history of exercise, activity levels, and aches and pains. Some displays showed information that would be tracked by self-report (eg, headaches). The goal with multiple examples for each display template was to encourage brainstorming, whereby the participant could feel comfortable giving gut reactions to specific ideas, while getting acquainted with deeper tracking concepts (eg, Textbox 2, #11) that they could then discuss.

Figure 3a. Body-based displays on (a) skin changes

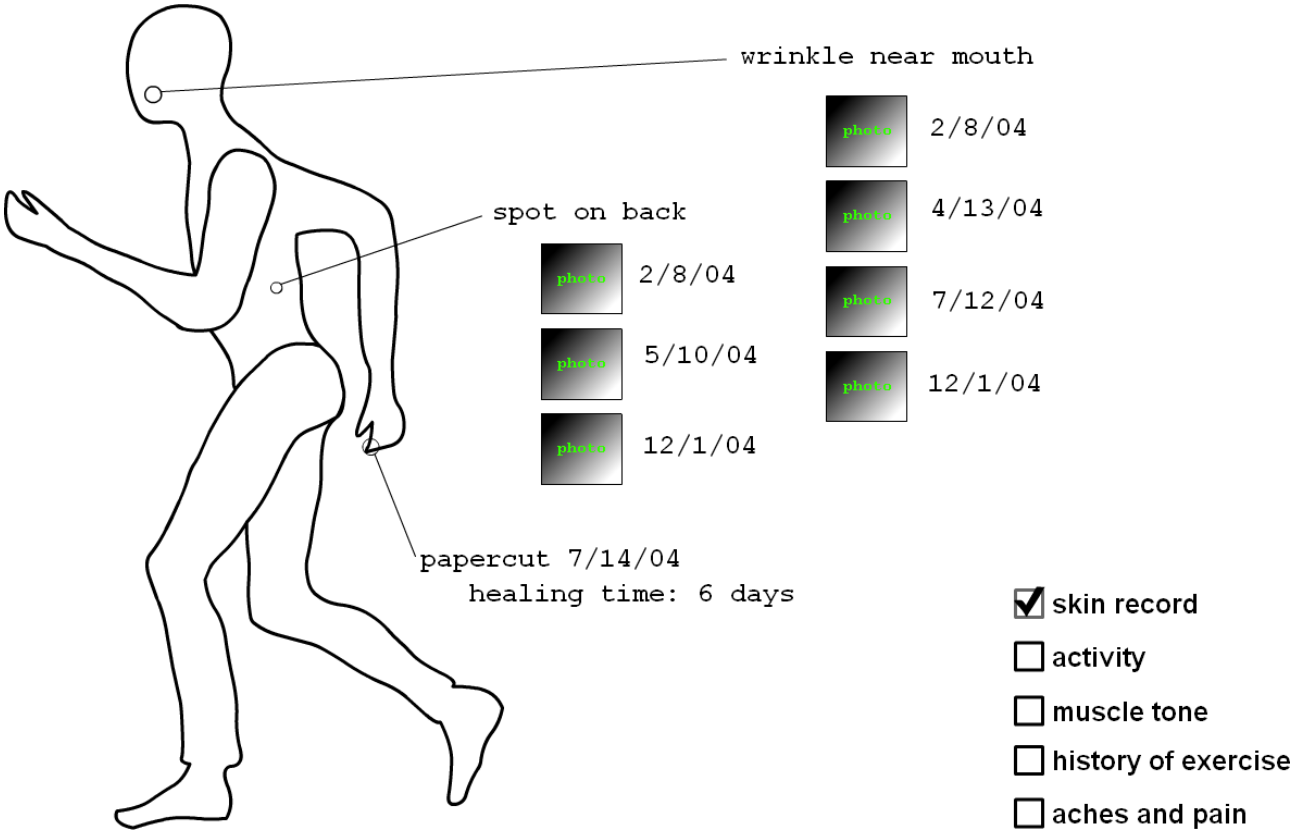
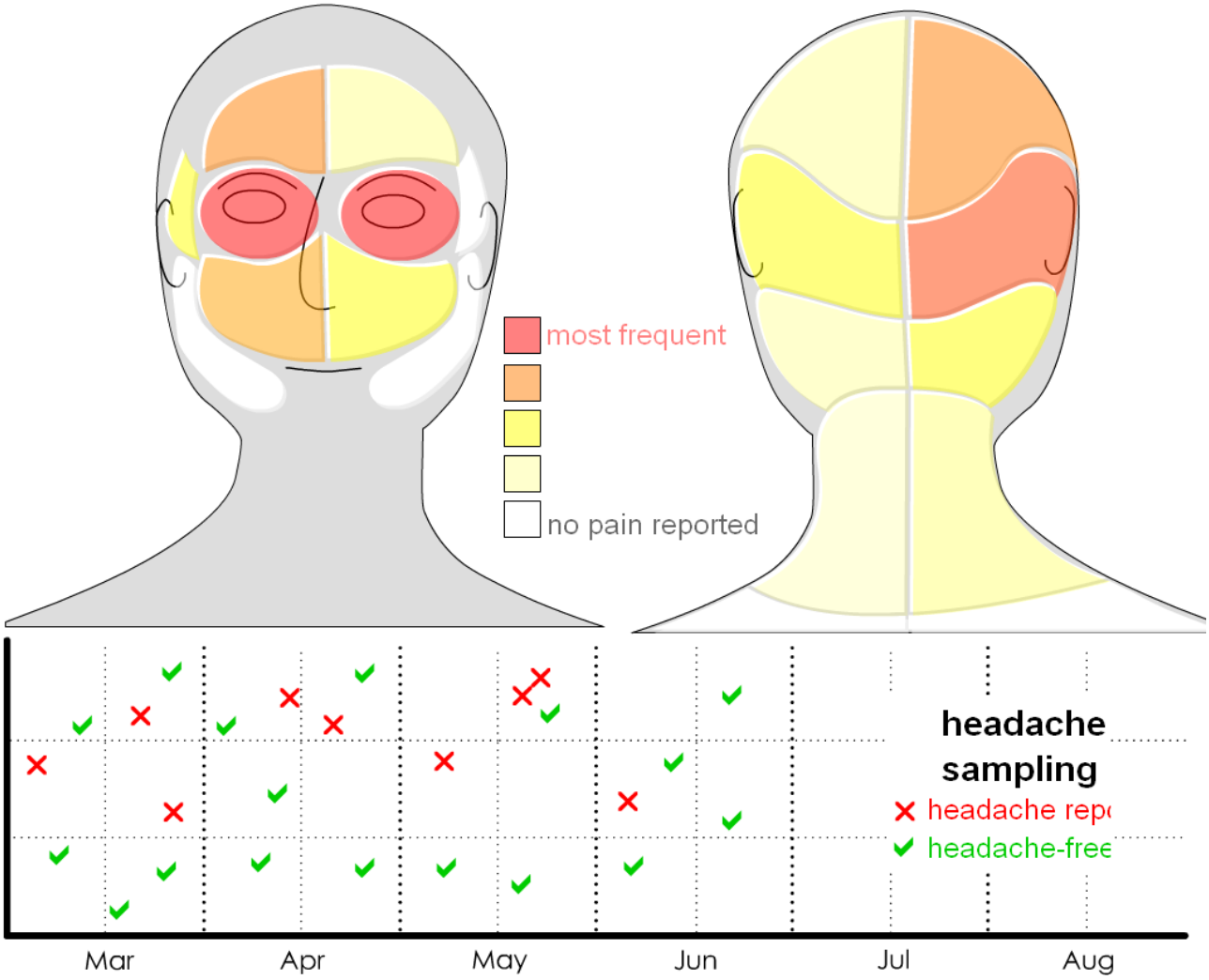


Figure 3b. and (b) headaches



Time Scale Displays Examples

More traditional graph-style displays were used to depict monitoring across multiple time scales (days, weeks, months, or years) (Figure 4). Y-axis units were not labeled to limit fixation on absolute data values that may not be salient for a particular participant. For example, time spent watching TV (Figure 4a) was represented with bars of varying heights, but there was no label of exact time. Variables of interest could be explored in isolation (eg, mood over time) or in relationship to other variables (eg, mood with respect to TV watching) (Textbox 2, #12)—the complexity of analysis depended on how many variables a participant checked in the screen display. On some displays, data could be compared with self-estimates (eg, your estimated TV watching) or group statistics (eg, average TV watching for your age group) (Textbox 2, #3). The years example (Figure 4b) was used to ask participants about variables that may be sensitive to major health changes.

Figure 4a. Time scale displays on (a) time spent watching TV (orange: TV watching; green: time spent outside; black dots: mood rating)

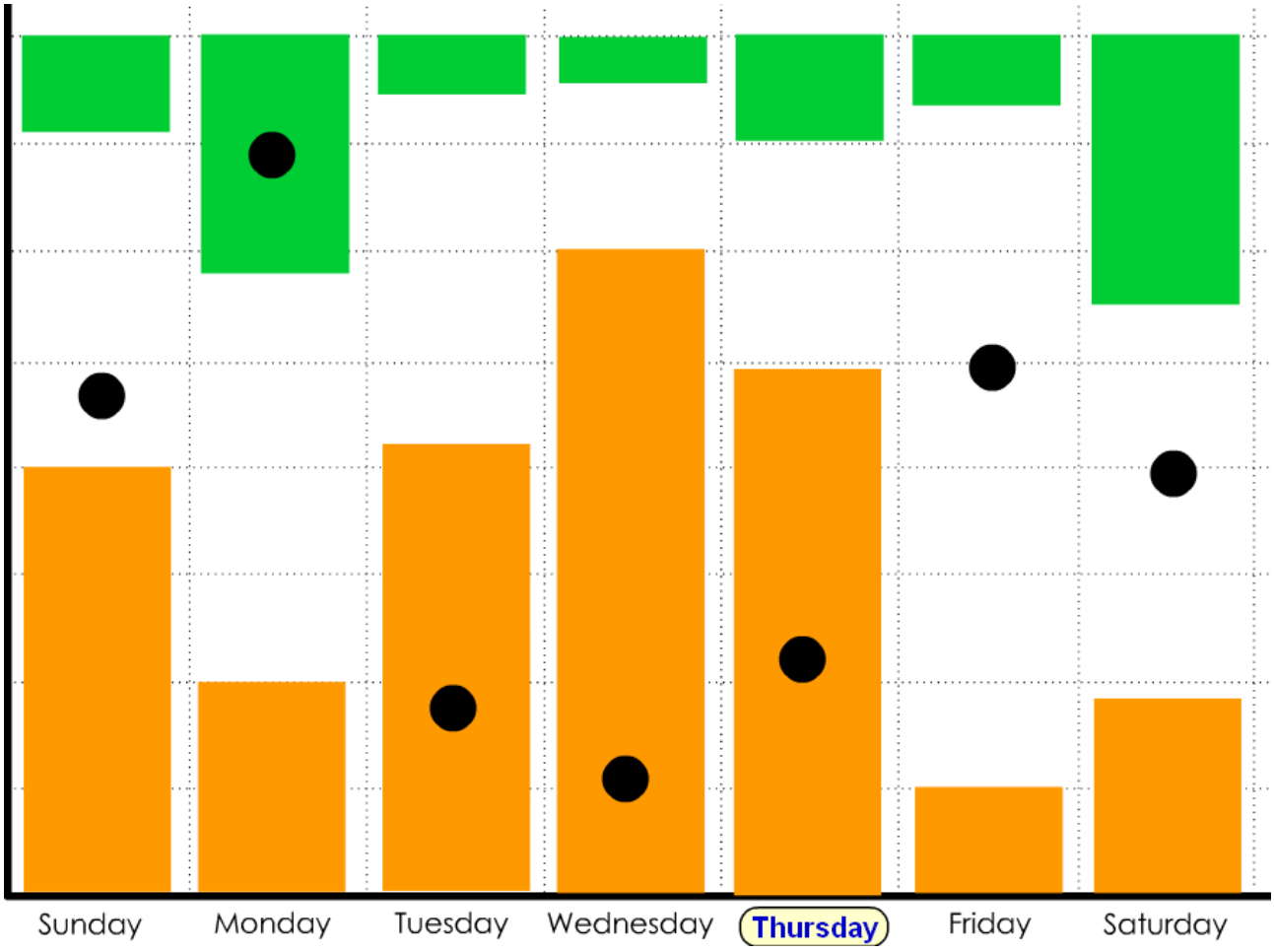
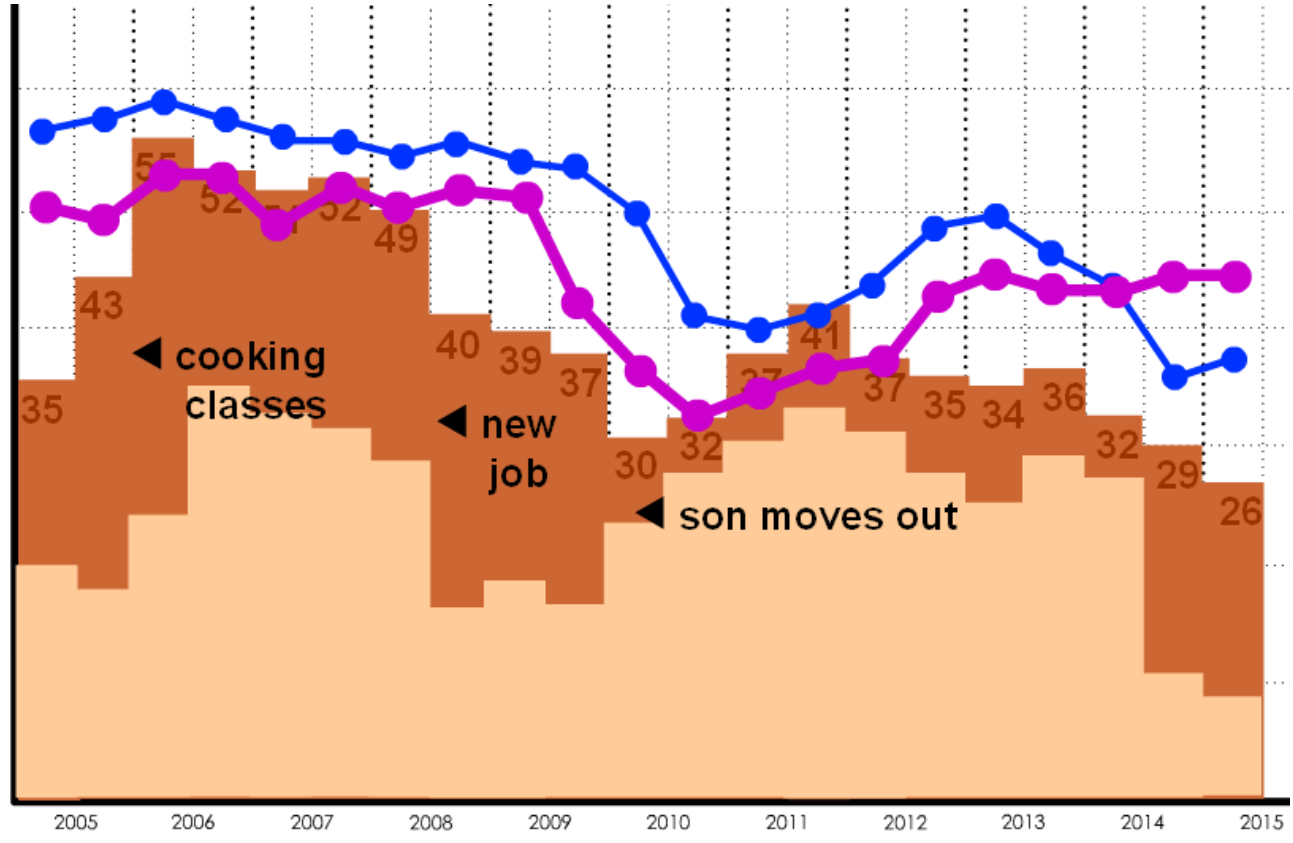


Figure 4b. and (b) variables that may be sensitive to major health changes



Interview Procedure with Health Professionals

We interviewed eight health professionals whom we identified through a working group on patient-centered health care at the Massachusetts General Hospital and through word-of-mouth. In contrast to the first set of experts from the preliminary interviews, many of these professionals treated healthy individuals and expressed a desire to shift to a coaching style doctor-patient relationship. They included a personal trainer, two social workers, three general practitioners, and two general practitioners who were cardiac specialists.

The interviews were 1-2 hours in duration. For all but one interview, the paper-based materials were employed; one participant responded to the electronic-based materials. The health professionals were shown 5-10 displays, starting with the weight displays, chosen because they depict a kind of patient-directed tracking with which health professionals may already have experience (ie, weight and pedometer steps).

Participants were instructed to imagine that the displayed data were collected in the patient's home and returned to the patient for personal reflection and conversations with a doctor or family member. They were reminded that the data in the displays do not necessarily reflect the capabilities of current technologies. For each example, they were asked to comment on how the information might be interpreted and applied by a patient, what it would be like to talk with a patient about the information, and what additional information would be beneficial to provide. They were encouraged to voice any concerns they might have about the collection and application of the data from their perspective as a health care provider.

Interview Procedure with Laypeople

We recruited 21 participants via postering, mailing lists, and word-of-mouth. All were from the United States. Participants volunteered their time without compensation for a protocol approved by our institutional human subjects review board. Included were 15 women and 6 men, ages 40 to 66 years (mean = 51; SD = 9). Although our target demographic was people in middle life, we additionally interviewed five individuals over age 70 for comparison. Three participants were close affiliates of one of the authors. Seven participants were from the local area of Cambridge/Boston, MA. The remaining participants were interviewed over the phone and represented 10 states.

The interviews were 1 hour in duration. In all interviews, the Web-based data displays were employed. Participants were shown 4-11 displays, typically starting with the journal example, chosen because it depicts a kind of tracking with which laypeople may already have experience (ie, diary or journal entries). They were instructed to imagine that the displayed data were collected in their home and returned to them for personal reflection. They were reminded that the data in the displays do not necessarily reflect the capabilities of current technologies. For each example, they were asked to identify content or qualities of interest or concern, how they might personally apply or use the information, and what additional information would be beneficial to have. The interviews were transcribed (masked transcripts with research notes are available in Multimedia Appendix 2).

Sorting Exercises with Laypeople

The displays were used to provoke discussion about health concerns that people wanted to investigate and the ways they could imagine conducting such investigations. We also wanted to quickly elicit specific ideas and rankings for concepts to track. We developed a list of 60 sample constructs that could be tracked over time using current or proposed tracking devices. These constructs represented multiple levels of inference or granularity and were selected to cover a diverse set of domains, including social interaction, cognition, physical activity, and physiology. Examples include “mood self-rating,” “blood pressure,” “tossing and turning,” and “time spent in the car” (see Multimedia Appendix 3 for the complete list).

After discussing the displays, the layperson participants in middle life (n = 21) spent 15 minutes doing the sorting exercises. The health professionals and five older adults also did a sorting exercise, but in a more open-ended manner, so for clarity, their results are not included here.

For interviews conducted face-to-face, participants were given a set of hand-written index cards to sort. For phone interviews, participants were directed to a Web page with a list of the constructs. Participants were asked to talk aloud while they sorted to indicate their choice (yes, no, or maybe) and comment about particular reasons for their decisions. Participants were asked to quickly sort the constructs according to whether they would personally want to track each over time. A second sorting exercise, where participants sorted with a particular investigative goal in mind, gave participants the opportunity to reconsider their initial gut reactions and express more focused applications for tracking. Participants were asked to select from a list of personal health areas that they might want to better understand and positively affect using tracking. The list included eating choices, family relationships, mental sharpness, mood, physical activity, and stress. Participants (n = 20) then sorted the set of constructs again, according to whether they would personally want to track each to understand their selected goal. One participant (under age 70) did not do the investigation sort due to time restrictions on the interview.

Results

We now review a few specific reactions from participants to the data displays and summarize the perceived benefits and limitations of health tracking from the perspective of the practitioners and laypeople. We also present variables of particular interest to the interviewees, as well as their specific ideas for applications of collected data.

Interviews with Health Professionals

In examining the body display (Figure 3), an interviewee who is a general practitioner first identified a benefit to the patient in having this type of data; he suggested, “Perhaps it could arm the patient so they can go in and more effectively communicate their worry [to a practitioner].”

Textbox 3. Longitudinal tracking benefits and concerns perceived by health professionals, summarized from the 8 preliminary interviews and the 8 materials-based interviews

Professional Perceived Benefits

Longitudinal data could be used to motivate and reward progress toward long-term goals (eg, exercise, nutrition); shift to life-goal assessments/recommendations, shift to "contracts" for change. (n = 6)

Longitudinal data can help doctors broach sensitive topics (eg, health of social relationships). (n = 4)

Longitudinal data can help doctors ask interesting questions and initiate a dialogue with the patient (n = 3); data can help patients communicate concerns. (n = 1)

Data collected outside the clinic is likely to be more representative of the patient's actual health. (n = 2)

Longitudinal data could be used to evaluate the success of interventions. (n = 2)

Longitudinal data may reveal that some constructs are more cyclic and context-based (e.g. mood corresponding to hormone cycles). (n = 1)

Longitudinal data could be used to detect the precursors of problem behaviors, such as eating disorders. (n = 1)

Longitudinal data can help the doctor make the most of a limited clinical visit. (n = 1)

Professional Perceived Limitations

Family physicians will not have the time or willingness to evaluate the data. (n = 4)

Most of what could be monitored either is not useful or is known already by the doctor, home nurse, caregiver, or patient. (n = 3)

Patients lack the ability to be self-aware about cognitive changes. Denial is too strong to overcome and, when serious, the cognitive impairments themselves make awareness difficult. (n = 3)

Patients may feel like their privacy or self-image is threatened. (n = 3)

Data will not represent more meaningful life pursuits (eg, becoming more self-aware, artistic advancement, enriched mental life). (n = 2)

Current generation of older adults is not interested in taking more responsibility for health. (n = 2)

Patients may become obsessed with data or may take a too narrow view, rather than approaching health holistically. (n = 2)

Data demonstrating decline will be depressing or disempowering. (n = 2)

If viewed over too long a time period, data will encourage a nonadaptive self-image. Instead of recognizing that one's current capacities are suited for the current context and meet basic life goals, one may focus too much on how one is not the person one used to be. (n = 1)

Patients may make snap judgments about contextual factors associated with health changes (correlation is not causation). (n = 1)

But he then began to imagine a counterproductive result, role playing the following interaction:

> *The patient who would bring this in, I would not look forward to seeing as a doctor…. [This kind of data] almost lets the patient live and dwell in something that you would rather have them just say—"I have these reoccurring headaches"—great, thank you, now how's the rest of your life going? It obscures rather than clarifies, in the conversation, they become too focused on the minute.*

Although this health professional strongly supported giving patients more control over their health through information, he worried about patients becoming obsessed with data and ignoring the larger issue of holistic wellness—a trend he has seen with medical professionals who over-rely on technology-generated data.

One of the interviewed social workers consults with her patients to identify resources and opportunities to establish healthy patterns. Reacting to a display graph depicting "kitchen cabinet access events" over time, she first noted that it would be useful in her work with patients with eating disorders. These patients often display food-search behavior as a precursor to destructive eating episodes. She envisioned using the data to draw patients into discussions about this behavior pattern and strategies for behavioral change. She then noted a second application: identifying "tea and toast ladies"—older adults who have over-simplified their diet following cognitive, physical, or emotional impairment.

The displays shifted interview foci from possible monitoring applications to personal scenarios, such as "What would it be like to have a patient bring me this data?" Although many of the presented ideas represented a departure from clinical metrics, they evoked a generally positive response and elicited more detailed feedback on using longitudinal data to develop individualized treatment plans and support patient-doctor communication. With the exception of one social worker who specializes in geriatrics and who was pessimistic that additional information or greater awareness would be relevant for late-stage health problems, the interviewed professionals thought that longitudinal tracking tools targeted to patients would be beneficial for the type of health practice they would like to have. However, several noted that the typical medical clinician would not be receptive. One participant responded, "A presently trained and configured family…doctor would think it was junk. They would go, 'Well, that's interesting, I got 9 more minutes.'"

An additional concern that appeared in multiple interviews was that the data should be presented in ways that do not force evaluation by comparison with norms, baselines, or externally determined goals or by presentation in a "report card" format. That said, some professionals indicated a desire to use data to motivate patients, one even arguing that the "report card" format was desirable.

We have summarized some of the health professional's key perceived benefits and limitations about longitudinal tracking in Textbox 3.

Interviews with Laypeople

At the beginning of the interview, participants were asked what they currently monitor or track. A man, age 40, who was working in a technology field, was a serious health tracker, manually recording multiple variables related to exercise, diet, and mood, including satisfaction ratings for leisure activities. Another man, age 40, working in a technology field, was a committed blogger, recording his intellectual process for work and maintaining a personal "kids" blog for his family. These participants were unusual in their existing dedication to tracking. However, all participants engaged in some type of monitoring (such as for weight, blood pressure, water drinking). Some kept journals (n = 9), tracked expenses (n = 2), kept annotated calendars (n = 7), and compiled lists of movies/books they watched/read/recommended as a backup for memory (n = 5). Several owned pedometers (n = 7), but no one was using them to track and review activity over time, citing reasons of forgetting and reluctance to exert the effort to record values.

We have summarized and categorized some of the laypeople's key perceived benefits and concerns about longitudinal tracking in Textbox 4. Here we describe a few specific reactions from participants to the data displays. For example, a woman, aged 66, who was working in an administrative position, initially responded to longitudinal tracking with disdain, asserting that she "already knew" the information it could provide and that anything she did not know would be better addressed by working with a professional. "It is interesting to me that you and the folks you are working with had the idea that people would like to have this information." After being introduced to the years example, as she expressed her reactions out loud, she gradually shifted perspective to consider the value of passive monitoring to establish a baseline and identify change. She did not want tracking as a tool for reflection and behavior change, but decided that background monitoring would be beneficial as she gets older. She did not want to use "it as a tool for understanding [her] life right now, but [to] monitor and anticipate change as [she is] aging, as an early warning to go to a doctor .... You've flossed every day for the last 12 years and in the last month, you've skipped 5 days, so what's going on?"

A 45-year-old man, who was a working professional, had re-purposed technologies to do some of his own tracking investigations. Prompted by a personal disagreement, he had recently reviewed years-old emails to gain perspective on the evolution of a relationship. He kept a personal calendar on his cell phone and noted, in reaction to one of the displays, that he is more conscious of commitments that are hard for him to keep (addressing finances) because he has a record of postponed appointments. He was strongly positive about an example that provided information about daily rhythms, believing that even without knowing what to look for, he would still find interesting patterns. "I think I might discover something…maybe correlate this with the journal to see how sleep patterns affect your overall mood or…amount of time you spend with other people affects your outlook." He worried though that having to describe his current mood states might amplify negative feelings (eg, saying you are sad makes you sadder). He was interested in how tracking could help him find more personal time. As a father, he could imagine using the group data to ground family discussions about issues such as bed time and together time, but was concerned that, in the end, the data might work to confirm and entrench opposing opinions. "So everyone will find the facts to support their interpretation of what's happening and the more facts you gather, the more it tends to just cement the positions of the people. That's been my experience."

A woman, aged 58, who was recently retired, wondered whether she could use longitudinal tracking to determine the amount of structure in her day that would give her the most satisfaction. She was interested in carrying out a variety of investigations, wanting to understand why she made qualitative judgments about a movie or a meal and what factors contributed to her mood. "It's like oh, today was really good—well, why was it good? Was it because of some really trivial compliment that you were given early in the morning, or is it because you accomplished something…. Was it just your blood sugar in the morning?" She identified eating habits as a behavior she would like to affect and expressed a desire for tools that would let her evaluate her personal degree of success with recommended interventions, such as drinking water to curb hunger. Additionally, she wanted data about her menopausal symptoms (eg, mood, memory problems) to persuade her reluctant physician to prescribe hormone therapy.

Overall, seven participants were reluctant to track, often stating that they "already know" sufficient information about their health and activities and preferred solutions (such as reminders) and expert-mediated care. Nevertheless, they each indicated that some tracking would be useful, either to establish a baseline and detect troubling changes or to address very directed investigations related to exercise or diet. Thirteen participants responded positively to multiple tracking ideas, particularly those related to behavior change and time management, but were somewhat conservative in proposing novel tracking investigations. Six participants were strongly positive, expressing comfort with exploring patterns and relationships in data, examining data on multiple time scales, and initiating involved or novel tracking investigations. Participants often qualified positive reactions by saying that the tracking idea would be just "fun" or interesting, but not necessarily valuable. They often initially focused on isolated and simple investigations (eg, frequency of headache pain), but with prompting, generated ideas about possible contributing contextual factors (eg, amount of sleep, shoes you are wearing) that when tracked could give them opportunities to problem solve. When participants strongly connected with a tracking idea, they often cited significant life events (eg, the 58-year-old's recent retirement and experience with cancer) as giving them a deeper appreciation for the

contextual factors related to general wellness. As the interview progressed, participants frequently identified behaviors or variables that were difficult to self-monitor (habits, behaviors done in combination with others, accumulation of subtle changes, personality variables, too familiar routines or factors) and were therefore ideal candidates for technology-assisted tracking.

Textbox 4. Longitudinal tracking benefits and concerns perceived by laypeople, summarized from interview transcripts

Layperson Perceived Benefits

Tracking may be useful for…

Supporting behavior change: Motivating/enabling change (n = 7); understanding cause and effect (n = 5); reinforcement, forcing to face issue (n = 5); evaluating success of intervention (n = 4); time management (n = 4); inspiring problem solving (n = 3); achieving "optimal state" or "peak condition" (n = 3); permitting trade-offs (since I've been good all week, it's okay to cheat) (n = 2); giving a sense of accomplishment (n = 2); identification of need for replacement activity (I was more active when I had kids, I had a more regular routine when I lived by the beach) (n = 1); measuring progress (n = 1)

Making patterns more evident: Understanding habits, traits, behaviors (n = 6); identifying factors that affect you positively or negatively (n = 5); identifying sources of distraction (n = 3); rethinking major life patterns (n = 4); identifying valuable improvisations (n = 1); understanding consequences of life-work balance (n = 2)

Monitoring health: Providing material for conversation with doctors or family (n = 11); determining seriousness of problem (n = 4); identifying troubling changes (n = 4); in place or available for later problems (n = 3); establishing a baseline for later comparison (n = 2); understanding nature of change (gradual versus sudden, onset) (n = 2); early detection, chance to cure, make plans for compensation (n = 1)

Challenging or validating beliefs: Getting perspective (n = 6); giving a sense of control (n = 5); promoting a positive self-image (n = 5); validation of subjective feelings (n = 1); peace of mind (monitoring parents) (n = 1)

Providing a record of events: Jogging/strengthening memory (n = 6); backup for memory (n = 6)

Providing entertainment/Supporting social interaction: Feeding curiosity, fun, interesting (n = 8); serving other or dual goals (reminders, communication) (n = 4); addressing physical appearance (n = 1)

Layperson Perceived Limitations

I wouldn't want to track (a variable or in general) because tracking would…

Not apply to me: (eg, smoking, alcohol drinking, pets) (n = 26)

Not provide new information: (ie, "I already know this") (n = 22)

Not provide valuable information: Something/someone already takes care of that (n = 6); unimportant or irrelevant (n = 5); just want a solution, not more information (n = 3); no change needed—functioning okay even though impaired (n = 2)

Provide too much information (information overload): Too focused on minute details, missing the big picture (n = 8)

Threaten self-image: Lead to denial/feeling threatened (n = 9); would feel criticized (n = 5); don't want to think about (eg, hate exercising, don't want to spend time thinking about exercising) (n = 2); lead to uncomfortable sense of competition (n = 1)

Not provide actionable information: Concerned about becoming discouraged or depressed (n = 7); frustration if don't know what to do or can't do anything (n = 7)

Lead to social conflict: Forcing involvement between family members (n = 5); fueling conflict and entrenched opinions (n = 3); other household members would not want to participate (n = 3)

Promote obsessive or unhealthy reactions: Living in the past (n = 3); will make too self-focused (n = 3); becoming obsessed (n = 3); becoming dependant and not thinking for self (n = 1); may use as an excuse not to change behavior (n = 1); will ignore after a while (n = 1); will amplify negative feelings (n = 1)

Force too much structure: Approaching life too analytically (n = 6); negative feelings about technology aspect (n = 3); reducing things of personal importance to a number (n = 1)

Not be suitable for particular activity or behavior: Very little or no change so tracking not needed (n = 2); behavior too erratic to be recognizable pattern (n = 2); activity is so routine, no need to track (n = 1); aspect of health is already bad (eg, night vision) (n = 1)

Be too complicated, error-prone, or disruptive: Need expert to interpret data (n = 8); data won't be accurate (because, eg, being observed, combined activity of family, algorithm may be biased) (n = 5); too much effort or time required (n = 6); privacy concerns (n = 2); data collection will be disruptive (n= 1)

Sorting Exercises with Laypeople

Participants frequently expressed surprise at how many constructs they selected to track. On average, participants indicated that they wanted to track more than a third of the constructs, answering "Yes, I want to track" for an average of 28.6 constructs out of 60 during the general sort.

The constructs selected most frequently for both exercises are listed in Table 1 (see Multimedia Appendix 4 for the complete table). While more traditional clinical metrics (blood pressure,

heart rate, blood sugar, hormone levels) are represented, there are also many constructs related to the quality of activity (multitasking, variation from routine), mental states (short-term memory, ability to concentrate, laughing), and behaviors (snacking, idle time, bed time).

Table 1. Constructs selected by at least 60% of the layperson participants in the general and investigation sorting exercises

Most Frequently Selected Constructs			
General Sort (n = 21)		**Investigation Sort (n = 20)**	
Construct	**Percentage**	**Construct**	**Percentage**
Correspondence with friends and family	81	Time at which you go to sleep	80
Heart rate	76	Ability to concentrate	75
Muscle tone	76	Idle time	75
Short term memory	71	Hormone levels/cycles	70
Pitch perception (hearing)	71	Heart rate	65
Time at which you go to sleep	71	Commitments	60
Use of space	71	Variation from routine	60
Ability to concentrate	67		
Hormone levels/cycles	67		
Laughing	67		
Multitasking	67		
Snacking	67		
Blood pressure	62		
Blood sugar (glucose)	62		
Commitments	62		
Posture	62		
Variation from routine	62		

While sorting, participants sometimes expressed personal reasons for selecting a construct. For example, one participant selected *alcohol drinking* not out of concern for excess, but because he was interested in keeping a record of his wine preferences. Another participant, in selecting *alcohol drinking* mentioned the recent purported health benefits of a glass of red wine per day, perhaps thinking of tracking as a method of recasting a pleasurable activity (drinking wine) as a health-directed activity. A third participant chose *alcohol drinking* expressing concern that her husband was drinking too much, perhaps looking to make him more aware of his own patterns if they tracked consumption as a couple.

The second sorting exercise, the “investigation sort,” offered participants an opportunity to confirm or modify their tracking selections according to a more concrete tracking investigation. Physical activity and stress were the most frequently selected investigations. On average, participants answered “Yes, I want to track” on 20.8 constructs out of 60. Although participants chose fewer constructs on average for an investigation, 18 out of 20 participants added constructs that they had previously categorized as “No, I don't want to track” on the general sorting exercise. On average, participants reassigned 5 constructs from “No” to “Yes” when asked to think about investigating a particular area of their lives. Examples of these constructs include idle time, awareness of time, ability to concentrate, and impulsiveness.

We now describe an example of how re-sorting using the investigation sort exercise inspired a participant to express a desire to track a variable that she would ordinarily avoid.

A semi-retired woman, aged 59, who was working as a part-time chef, first sorted the constructs according to whether she would like to track them over time. She put *idle time* in the “no” category, explaining that she has a tendency to think that she has to “stay busy” and would feel uncomfortable examining times when she is not. After completing the general sort, she was asked to select an aspect of her life that she wanted to learn more about or change. She chose *stress* and sorted again. There were a few constructs that she selected in the general sort, such as *use of space*, that didn't seem to apply for this focused investigation, so she put them in the “no” category. However, there were several that she had previously rejected, including *clothing choices*, *night vision*, *trips to the grocery*, and *idle time*, that she chose for this re-sort. She was surprised that the more focused inquiry altered her choices. She explained why she changed her mind about tracking *idle time*: “[When] I'm in a down mood, idle time really is disturbing to me.” She went on to explain that her reaction to idle time was therefore a useful indicator of degree of stress.

Discussion

Here we summarize some of our impressions, which we believe would be helpful for designers of health tracking systems. We qualify this discussion with the caveat that cultural differences and differences in medical care between countries could lead to different results in non-US populations using the same data displays.

Our participants envisioned conducting customized, short-term health investigations. People in middle life are often trying to address complex personal health issues, such as regular fitness routines and complex diets; sometimes these goals seem to dominate their lives. They are often dealing with radical changes to their lifestyle, such as divorce and retirement. Most participants wanted to be able to conduct short investigations, either to address a desired behavior change or to identify factors influencing their health and behavior. Typically these investigations would not start from scratch. The person has a theory about a factor that may be influencing how she or he feels and would appreciate tools that help isolate the other factors that influence the condition. Daily rhythms (including variation from routine, sleep/wake cycles, and time spent inside/outside), habitual behaviors (eating, posture), and physical health and activity were of consistent interest. Many people were also interested in mood and social interaction.

Participants frequently expressed an interest in monitoring patterns that would probably not be considered "health"-related by medical professionals. For example, some expressed a desire to be able to examine personal living patterns and variables such as time management, life-work balance, and social relationships, all variables that change as a result of life stage and other factors. These may be related to health indirectly, but they were typically not mentioned by the first set of health professionals as factors that should be tracked.

In direct contrast with concerns expressed by professionals, most laypeople wanted data about their health and behavior even if it invited negative self-evaluation (eg, declines in short-term memory, excessive time spent watching TV). They felt this information would motivate them to change a problematic behavior or help them determine how to get preventive care and compensate for health and performance changes. Another concern expressed by health professionals was evaluation against norms. Most laypeople interviewed were prepared to evaluate data according to their own expectations and goals and without a need for norms or expert evaluation (with some exceptions, such as skin changes).

Concerns about intrusiveness primarily occurred when examining the communication of data between family members. Participants were reluctant to "force involvement" by sharing data about their own activities and health with family members. Two parents with young children and adolescents suggested that there were some data they did not need themselves but that their children would benefit from (eg, time spent on morning routines). Participants suggested that family members may not be willing to share in the tracking investigations or might use data to fuel conflict. It was also evident, however, that participants were interested in leveraging social relationships to address health issues. Privacy issues also emerged when participants considered monitoring their behavior at work.

Constructs that had a duration component (eg, time spent cooking) were considered stressful or overly structured by many people. Data on short time scales were often seen as redundant to memory or provoked feelings of self-consciousness. Some participants found the concept of tracking over years difficult to grasp and struggled to identify variables they would want to track; others expressed that the long time scales might help them gain perspective on lifestyle and re-assess life goals.

Some participants felt that data collection or intervention tools that explicitly interrupt the user may be perceived as reinforcing negative states (eg, having to say you are sad may make you sadder). An exception to this was a tool that might help the user identify and get perspective on rising anger.

Finally, there was a tendency to focus on reprimand—data that tells them what they are doing wrong (eg, "how much time I'm wasting," "how I'm not calling my family enough") rather than on how data can help them solve a problem or reinforce positive self-evaluation (eg, "I *am* spending enough time calling my family"). This may reflect a bias ingrained by our medical system. Tests or professional medical advice are most often either ignorable if the news is positive (eg, no cancer detected) or they indicate bad news about a condition or a behavior. Put simply, no news is generally good news. A ubiquitous computing system that may collect and provide data to proactively help improve physical, social, and mental well-being is a concept that most people have not yet considered.

Effectiveness of Methodology

We developed the mock data display interview material as a response to the challenges of eliciting rich and personal responses to concepts for health monitoring technologies using traditional talk interview techniques. They were developed as a way to help participants respond as though they had experienced the technologies without being focused on the rough implementation details of currently available research prototypes. Because participants were offered concepts to react to and expand upon, rather than being asked to respond from a "blank slate," assumptions held by the researchers about the range and affordances of health monitoring data were undoubtedly conveyed and therefore impacted participant response. Overall, however, we found the use of the data displays to be helpful for eliciting a variety of health tracking concerns as well as ideas for potential applications of longitudinal health data from both practitioners and laypeople. We did encounter a few challenges. First, drawing out personal responses sometimes took some effort. Many participants initially stated reactions in terms of how it could be useful to people in general rather than to them personally and required prompting to focus on their own needs. Second, sometimes participants stated that a factor applied to them but subsequently expressed indifference about tracking and reflecting upon it. It was easy for participants to respond "Oh yes, I email" or "I don't watch much TV" rather than indicating whether they wanted to track the variable.

Participants frequently said that they were well aware of their behavioral, emotional, and performance patterns. We suspect that they may be overestimating their self-awareness given that retrospective recall tasks for common events is typically very poor [26]. A useful line of research would be to determine how accurately people can retrospectively recall and evaluate some of the factors they believe they are familiar with. For those where perception and reality are different, an interesting question is whether people should be made aware of the discrepancy.

As a group, our participants generated an impressive list of ideas for directed investigations (eg, "What is the level of tension in the house?", "How close am I to needing to consider back surgery?"). It was therefore notable when participants struggled with a display, sometimes focusing on isolated variables, rather than variable relationships, or overlooking examples of antecedent conditions or contextual factors associated with behaviors (eg, actions and events leading up to skipping exercise). With prompting, they often saw the value in widening a health or behavior-change investigation to consider many more variables in addition to the obvious ones. Their difficulty generalizing from the displays to their own situations may be attributable to limitations in the mock data displays; however, further research is warranted to discover ways to best uncover people's mental models and curiosities about their own behavior that could motivate the design of novel ubiquitous health care applications.

Recommendations

Ubiquitous computing technology may offer impressive new capabilities for home monitoring to support traditional clinical diagnosis and health assessment. However, due to concerns about cost, privacy invasion, and how end users would react to such monitoring, our interviews suggest that a fruitful direction for researchers interested in home health monitoring to pursue may be to develop technologies that allow for personalized home tracking investigations. Our interviews clearly show that while there is great variability in what factors about their life people would want to track, most people in our interviews did have concepts they would longitudinally track and self-monitor given appropriate technology. Our interviews also suggest that what people wish to track will change over time, based upon their age, life circumstances, interactions with friends and family, health status, and general curiosity. Based upon these interviews, we recommend that ubiquitous "monitoring" systems may be more readily adopted by end users if they are developed as tools for personalized, longitudinal self-investigation that primarily help end users, instead of or in addition to medical professionals, learn about the conditions and variables that impact their social, cognitive, and physical health.

We advocate research on ubiquitous computing longitudinal health monitoring systems that do the following: (1) collect information with as little interaction required from the end user as possible; (2) find appropriate times and places for users to reflect, review, and initiate investigations; (3) help people conduct short-term investigations about issues they are curious about, sometimes "channeling" fun or interesting tracking into deeper, long-term health investigation; (4) build upon the health investigations that people are already engaged in, supporting people as they try diets, train for events, or participate in exercise programs; (5) support nonhealth areas of interest, such as tracking spending, time management, home design, communication with friends and family, and memory support, in order to facilitate collection of health-related data; (6) help people explore relationships between variables that at first they might not consider (eg, mood and TV watching) and appreciate unexpected variability; and (7) help people focus on accomplishments and positive behaviors.

Acknowledgments

We thank our interview and study participants for their insightful and detailed feedback and the generous contribution of their time. This work was supported by Intel Corporation and the MIT House_n Consortium.

Conflicts of Interest

None declared.

All Multimedia Appendices are also available at: http://architecture.mit.edu/house_n/data/tracking/

Multimedia Appendix 1

Mock data displays [SWF (Shockwave Flash) file, 432 KB - jmir_v8i4e29_app1.swf]

Multimedia Appendix 2

Masked interview transcripts [PDF file, 572 KB - jmir_v8i4e29_app2.pdf]

Multimedia Appendix 3

Complete list of sorting constructs [HTML file, 40 KB - jmir_v8i4e29_app3.html]

Multimedia Appendix 4

Complete table of selected constructs [PDF file, 12 KB - jmir_v8i4e29_app4.pdf]

References

1. Mynatt ED, Essa I, Rogers W. Increasing the opportunities for aging in place. In: Proceedings of the Conference on Universal Usability; 2000:65-71.
2. Bratan T, Clarke M, Jones R, Larkworthy A, Paul R. Evaluation of the practical feasibility and acceptability of home monitoring in residential homes. J Telemed Telecare 2005;11 Suppl 1(5):29-31. [Medline: 16035985] [doi: 10.1258/1357633054461796]
3. Korhonen I, Iivainen T, Lappalainen R, Tuomisto T, Kööbi T, Pentikäinen V, et al. TERVA: system for long-term monitoring of wellness at home. Telemed J E Health 2001;7(1):61-72. [Medline: 21220146] [doi: 10.1089/153056201300093958]
4. Philipose M, Fishkin KP, Fox D, Kautz H, Patterson D, Perkowitz M. Guide: towards understanding daily life via auto-identification and statistical analysis. In: Ubihealth 2003: The 2nd International Workshop on Ubiquitous Computing for Pervasive Healthcare Applications 2003.
5. Matsouoka K. Smart house understanding human behaviors: who did what, where, and when. Proceedings of the 8th World Multi-Conference on Systems, Cybernetics, and Informatics 2004;3:181-185.
6. Haigh KZ, Kiff LM, Myers J, Guralnik V, Krichbaum K, Phelps J, et al. The Independent LifeStyle Assistant™ (I.L.S.A.): Lessons Learned. Report. Minneapolis, MN: Honeywell Laboratories; Dec 01, 2003.
7. Korhonen I, Paavilainen P, Särelä A. Application of ubiquitous computing technologies for support of independent living of the elderly in real life settings. In: UbiHealth 2003: The 2nd International Workshop on Ubiquitous Computing for Pervasive Healthcare Applications 2003.
8. Mihailidis A, Carmichael B, Boger J, Fernie G. An intelligent environment to support aging-in-place, safety, and independence of older adults with dementia. In: UbiHealth 2003: The 2nd International Workshop on Ubiquitous Computing for Pervasive Healthcare Applications 2003.
9. Barger T, Brown D, Alwan M. Health status monitoring through analysis of behavioral patterns. In: Proceedings of the 8th Congress of the Italian Association for Artificial Intelligence (AI*IA) on Ambient Intelligence. New York: Springer-Verlag; 2003.
10. Wilson DH, Atkeson C. Simultaneous Tracking & Activity Recognition (STAR) Using Many Anonymous, Binary Sensors. In: Proceedings of PERVASIVE 2005. New York: Springer-Verlag; 2005.
11. Alwan M, Dalal S, Mack D, Kell SW, Turner B, Leachtenauer J, et al. Impact of monitoring technology in assisted living: outcome pilot. IEEE Trans Inf Technol Biomed 2006 Jan;10(1):192-198. [Medline: 102762135] [doi: 10.1109/TITB.2005.855552]
12. Lukowicz P, Kirstein T, Tröster G. Wearable systems for health care applications. Methods Inf Med 2004;43(3):232-238. [Medline: 15227552]
13. Lymberis A. Research and development of smart wearable health applications: the challenge ahead. Stud Health Technol Inform 2004;108:155-161. [Medline: 15718642]
14. Jimison HB, Pavel M, McKanna J, Pavel J. Home monitoring of computer interactions for the early detection of cognitive decline. In: Proceedings of the IEEE Engineering in Medicine and Biology Conference 2004.
15. Doughty K, Cameron K, Garner P. Three generations of telecare of the elderly. J Telemed Telecare 1996;2(2):71-80. [Medline: 98042635] [doi: 10.1258/1357633961929826]
16. Abidi SS. TIDE: an intelligent home-based healthcare information & diagnostic environment. Stud Health Technol Inform 1999;68:720-725. [Medline: 20109711]
17. Zhang D, Yu Z, Chin CY. Context-aware infrastructure for personalized healthcare. Stud Health Technol Inform 2005;117:154-163. [Medline: 16282665]
18. Kunze C, Grossmann U, Stork W, Müller-Glaser KD. Application of ubiquitous computing in personal health monitoring systems. Biomed Tech (Berl) 2002;47 Suppl 1 Pt 1:360-362. [Medline: 22340026]
19. Morris M, Intille SS, Beaudin JS. Embedded assessment: overcoming barriers to early detection with pervasive computing. In: Gellersen HW, Want R, Schmidt A, editors. Proceedings of PERVASIVE 2005. New York: Springer-Verlag; 2005:333-346.
20. Baos V, Ester F, Castellanos A, Nocea G, Caloto MT, Gerth WC, I-Max Study Group. Use of a structured migraine diary improves patient and physician communication about migraine disability and treatment outcomes. Int J Clin Pract 2005 Mar;59(3):281-286. [Medline: 15857323] [doi: 10.1111/j.1742-1241.2005.00469.x]
21. Richardson CR, Brown BB, Foley S, Dial KS, Lowery JC. Feasibility of adding enhanced pedometer feedback to nutritional counseling for weight loss. J Med Internet Res 2005;7(5):e56 [FREE Full text] [Medline: 16403720] [doi: 10.2196/jmir.7.5.e56]
22. Anhøj J, Nielsen L. Quantitative and qualitative usage data of an Internet-based asthma monitoring tool. J Med Internet Res 2004 Sep 3;6(3):e23 [FREE Full text] [Medline: 15471749] [doi: 10.2196/jmir.6.3.e23]

23. Anhøj J, Jensen AH. Using the internet for life style changes in diet and physical activity: a feasibility study. J Med Internet Res 2004 Sep 8;6(3):e28 [FREE Full text] [Medline: 15471754] [doi: 10.2196/jmir.6.3.e28]
24. Denton IC. Will patients use electronic personal health records? Responses from a real-life experience. J Healthc Inf Manag 2001;15(3):251-259. [Medline: 21520717]
25. Tang PC, Ash JS, Bates DW, Overhage JM, Sands DZ. Personal health records: definitions, benefits, and strategies for overcoming barriers to adoption. J Am Med Inform Assoc 2006;13(2):121-126. [Medline: 16357345] [doi: 10.1197/jamia.M2025]
26. Gorin AA, Stone AA. Recall biases and cognitive errors in retrospective self reports: a call for momentary assessments. In: Baum A, Revenson T, Singer J, editors. Handbook of health psychology. Mahwah, NJ: Erlbaum; 2001:405-413.

submitted 26.05.06; peer-reviewed by W Rogers; comments to author 06.08.06; revised version received 17.11.06; accepted 18.11.06; published 07.12.06.

Please cite as:
Beaudin JS, Intille SS, Morris ME
To Track or Not to Track: User Reactions to Concepts in Longitudinal Health Monitoring
J Med Internet Res 2006;8(4):e29
URL: http://www.jmir.org/2006/4/e29/
doi:10.2196/jmir.8.4.e29

Original Paper

Evaluation of Spoken Dialogue Technology for Real-Time Health Data Collection

Esther Levin[1], PhD; Alex Levin[2], BSc

[1]Department of Computer Science, City College of New York, New York, NY, USA
[2]Spacegate Inc, Livingston, NJ, USA

Corresponding Author:
Esther Levin, PhD
Department of Computer Science
City College of New York
Computer Science, NAC Bldg, Rm 7-312
138th Street and Convent Avenue
New York, NY 10031
USA
Phone: +1 973 568 5843
Fax: +1 814 253 0608
Email: esther@cs.ccny.cuny.edu

Abstract

Background: A real-time assessment of patients' experiences is an important methodology for studies in health care, quality of life, behavioral sciences, and new drug and treatment development. Ecological momentary assessment is a methodology that allows for real-time assessment of experience and behavior in a subject's natural environment. Recently, electronic data collection techniques have been introduced, including systems utilizing interactive voice response.

Objective: The objective of this project was evaluation of spoken dialogue methodology for real-time data collection of information from patients for health, behavioral, and lifestyle studies and monitoring. While the management of the data collection process was Internet-based, this additional eHealth communication channel was based on over-the-phone natural language conversation with a dialogue system utilizing automated speech recognition technology. For this study we implemented a dialogue system for patients' assessment and monitoring of chronic pain.

Methods: Experimental evaluation of usability of the Pain Monitoring Voice Diary was performed with 24 volunteers. The volunteers were asked to contribute 10 sessions with the system over a period of 2 weeks; in practice, the number of sessions per subject ranged from 1 to 20. The subjects were asked to either relate to pain episodes in their past while answering the system's questions, or use as a guidance one of nine provided medical scenarios compiled by a pain specialist, ranging from migraines and back pain to post-surgical pain (knee injury) and cancer- and chemotherapy-related afflictions.

Results: From 24 volunteers, we collected a total of 177 dialogue sessions: 171 sessions were completed, while the caller hung up in the other 6 sessions. There were a total of 2437 dialogue turns, where a dialogue turn corresponds to one system prompt and one user utterance. The data capture rate, measuring the percentage of slots filled automatically, was 98%, while the other 2% were flagged for transcription. Among the utterances sent to transcription, where the user had opted for the "none of those" option, 70% corresponded to the "type of pain" slot, 20% to the "symptoms" slot, and 10% to the "body part" slot, indicating that those are the grammars with the highest out-of-vocabulary rate.

Conclusions: The results of this feasibility study indicated that desired accuracy of data can be achieved with a high degree of automation (98% in the study) and that the users were indeed capable of utilizing the flexible interface, the sessions becoming more and more efficient as users' experience increased, both in terms of session duration and avoidance of troublesome dialogue situations.

(J Med Internet Res 2006;8(4):e30) doi:10.2196/jmir.8.4.e30

KEYWORDS

Human factors; ecological momentary assessment; data collection; voice recognition

Introduction

Use of questionnaires is an essential method of data collection, especially in studies of health care, quality of life, behavioral sciences, and new drug and treatment development. A real-time assessment of experience and behavior in the patient's natural environment is an important parameter that provides feedback and input to the health professional, researcher, or pharmaceutical company about the effects of treatment and/or the patient's quality of life. Very often the research or study findings are based significantly or completely on questionnaire responses. While designing valid questionnaires is an art, the tools and methods of data collection are not less important and often can influence the research outcome. Ecological momentary assessment (EMA) is a methodology that allows for real-time assessment of experience and behavior in a subject's natural environment. The methodology has evolved from the behavioral sciences [1-6] and enables the gathering of meta-data on patient compliance, as well as the measurement and improvement of compliance.

Traditional EMA data collection methods vary from paper-based diaries and reports to video/audio recordings and to human observation. However, doubts have been cast on the validity of the data collected through paper-based methods of self-report, notably from a recent study that demonstrated that most of the paper diary entries by patients (79%) were falsified [7]. Recently, electronic data collection techniques have been introduced, including personal digital assistants (PDAs), the Internet [8], and cellular phones utilizing interactive voice response (IVR). These methodologies enable collection of meta-data on the respondent's compliance and use of such data to measure and improve compliance. While PDAs and Internet collection methods demonstrate clear advantages over paper forms, they also have certain limitations, including use of uncommon devices requiring participant training, limited availability, and extensive data management and programming costs. In addition, PDA methods require in-person contact to download data and change batteries. In IVR-based data collection participants use phones to call the system and answer the question posed by the system by pushing keys on the phone. Collins et al have reported on the compatibility of phone-based IVR data collection and its improvement over paper-based and PDA-based systems [9].

Spoken language dialogue system for data collection [10-13] is based on automatic speech recognition and, similarly to IVR, is a phone-based approach. It extends the IVR approach by allowing the subject to communicate with the system using a natural spoken language as an input modality. A challenge such a system faces is maintaining adequately high accuracy of the captured data while guaranteeing a satisfactory user experience. In particular, since the subjects conduct dialogues with the system on a regular basis, adequate dialogue design should provide a flexible level of user support to accommodate both novice callers and experienced callers: for the experienced caller, the system needs to provide short and effective call flow without making the caller hear long and tedious prompts; for the novice caller, the system needs to provide enough information and help to guarantee question understanding and successful session completion.

In this paper we describe a design for spoken language dialogue that takes into account the specificities of the data collection task. We first present spoken dialogue technology and its potential applicability to data capture task. We then describe a Pain Monitoring Diary dialogue design and how it addresses specific requirements of data capture task. We then report the results of a feasibility study.

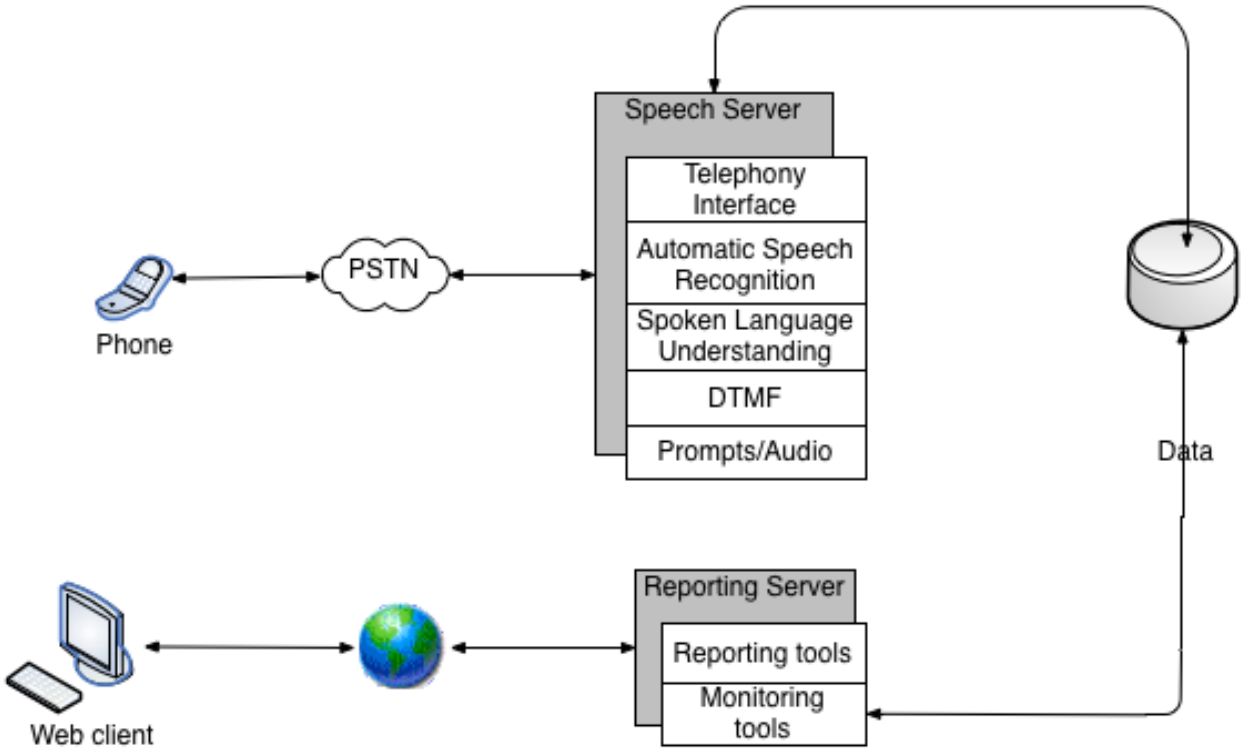
Methods

Data Capture via Spoken Dialogue System

Figure 1 presents the block diagram of a spoken dialogue system. The system has two interfaces: the first for the subjects and the second for the health care providers or researchers gathering the data. The subject communicates with the system through phone dialogue. A recording of such an interaction is included in Multimedia Appendix 1 and its transcription is shown in Textbox 1.

Each data collection session is scheduled according to a protocol and is initiated either by the system or by a subject. During the phone session, the utterances spoken by the subject are captured by the speech server. The speech server invokes automated speech recognition (ASR) and spoken language understanding modules to convert spoken utterances to textual semantic representation of subject's inputs and then stores them in the database. The reporting server can then use this data in real time to generate a report about the session that can be viewed by a health care provider or a researcher several seconds after the session ends. The reporting server works in a push or a pull mode, or both. In the pull mode, the reports can be viewed over the Internet on demand; in the push mode, the reporting server sends notifications and alerts (via email, cell phone, or pager) to the health care provider about the availability of the new report. Table 1 shows a report automatically generated by the system at the end of the session presented in Textbox 1.

Figure 1. Block diagram of the data collection system; PSTN: public switched telephone network; DTMF: dual tone multi-frequency



Textbox 1. Transcription of a sample dialogue

System: Your PIN, please.

Caller: ****

System: OK, *****, let's begin. Are you in pain now?

Caller: Yes.

System: Where does it hurt? [pause] For example, your head, stomach...

Caller: My left shoulder.

System: Was that your left shoulder?

Caller: Yes.

System: Please rate your pain on a scale of 0 to 10, with 0 being no pain and 10 being the worst possible pain.

Caller: It is 7.

System: Was that 7?

Caller: Yeah.

System: How does the pain feel?

Caller: I don't know.

System: Okay, here is the help information. At this point, I need to find out the type of your pain. Please choose carefully a word from the following list that best describes your pain and just say it. If nothing matches, please say "none of those." Here is the list: aching [pause], burning [pause], dull…

Caller: It feels sharp.

System: Was that "sharp?"

Caller: Yeah.

System: Is the pain constant or does it come and go?

Caller: Comes and goes.

System: Do you have any other symptoms?

Caller: Yeah.

System: Please tell me your most disturbing symptom.

Caller: I have fever.

System: I heard "fever." Is that right?

Caller: Yes.

System: Would you like to leave a message for the staff?

Caller: No.

System: Thanks for your cooperation. I've got all the information that I need. I'll call you again in 4 hours. Goodbye.

There are at least four advantages of using spoken dialogue technology for data collection compared to traditional methods using pen and pencil. First, speech is a natural modality of interaction for humans, and the input device—the phone—is user friendly and ubiquitous and no special training is required for its use (as opposed to PDAs or computers). Second, compliance is monitored automatically: the calls can be initiated by the system following a prescribed protocol, and the system can report any noncompliance to the trial administrator in real time. Third, spoken automated dialogue reaches much beyond voice-enabled static paper questionnaires: possible answers are not limited by the number of check-boxes to fit on a piece of paper; question selection can be done dynamically based on previous answers; and personalization of both content and style based on the patient's history is possible. Fourth, the ability to transform the captured data into real-time reports and further interface the information with other clinical or back-office systems and databases provides an unparalleled opportunity to enhance patient feedback and monitoring. Overall, an ASR-based system offers the health care provider an extensive and practical tool to facilitate efficient and convenient patient communications, which saves time while increasing quality of care.

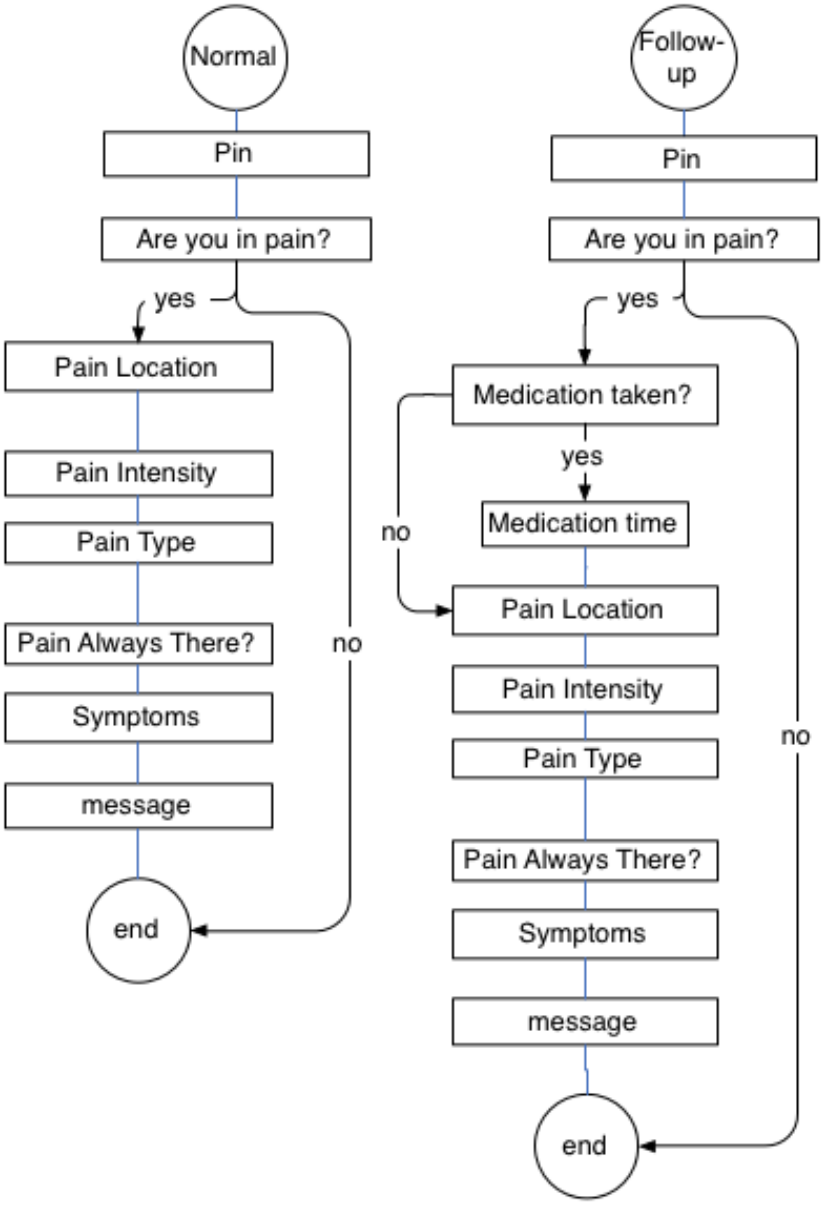
Table 1. Automatically generated report for the dialogue in Textbox 1

	Captured Value	Confirmed (yes/no)	Confidence Score
PIN	****	no	66
Are you in pain?	yes	no	80
Pain location	left shoulder	yes	86
Pain intensity	7	yes	88
Pain type	sharp	yes	88
Pain constant?	pain comes and goes	no	47
Symptoms	fever	yes	86
Message	none	no	78

Pain Monitoring Diary Application

For this study, we implemented a dialogue system for chronic pain patients' assessment and monitoring. Pain assessment is an application for which well-established standard questionnaires [14,15] are available, and the vocabulary for potential answers can be established from the medical literature. Figure 2 shows dialogue flow for the Pain Monitoring Diary. The dialogue flow is represented as a series of dialogue units, where each unit comprises several caller-system exchanges designed to elicit one piece of information from the caller to fill a slot in the session report. A slot is, for example, pain location, pain intensity, pain type. Please note the correspondence between the dialogue units in Figure 2 and the dialogue session (Textbox 1) and its report (Table 1).

Figure 2. Schematic illustration of call flow for the Pain Monitoring Diary in normal and follow-up modes



Dialogue Design

The characteristics and requirements of data capture tasks are different than those for other applications of spoken dialogue technology [16]. Successful dialogue design needs to take the following specificities of this task into account.

First, data validity, accuracy, and integrity in this application are very important since the penalty for an erroneously filed final session report can be very high. Since the ASR technology is not perfect, the design has to take into account the possibility of speech recognition errors and improve the overall accuracy using dialogue actions such as re-prompts, confirmations, error handling, and, if necessary, recording and flagging the unrecognized utterances for later transcription.

Second, subjects call the system repeatedly according to the study protocol and identify themselves at the beginning of the session. This provides an opportunity to use the knowledge accumulated across sessions for personalization.

Third, the system should accommodate both novice callers (in the beginning of the trial) and experienced callers (those who completed several sessions). For the experienced caller, the system needs to provide a short and effective call flow, without making the caller hear long and tedious prompts. For the novice caller, the system needs to provide enough information and help to guarantee question understanding and successful session completion.

Fourth, the subjects participating in data collection are enrolled through a personal face-to-face interview during which they receive relevant information about the trial and guidance on the process of data collection. In the same opportunity, the participants can receive some training, explanation, and possibly a demonstration on how to use the spoken dialogue system.

Based on these considerations, our goals in the design of the dialogue system were controlling the accuracy of the captured data while providing a flexible and adequate level of user

support allowing efficient communication for experienced users and sufficient support for novices.

Controlling Captured Data Accuracy

We designed the system to take into account the known limitations of ASR technology and to be able to ensure overall high accuracy of data capture and session completion rate. In general, the most important parameter that determines the accuracy of speech recognition is the size and the complexity of the grammar that is being used for recognition of the current utterance. The grammar in speech recognition describes (as text) the set of all the possible sentences that can be recognized by the system. For example, the simplest grammar that can be used in yes/no recognition will contain only two sentences: {"yes," "no"}. During speech recognition, current user utterance is matched to every sentence in the grammar, and the recognition result is given as the best matching sentence within the grammar, together with a recognition score that measures the quality of the match. If the quality of the match is not good enough, the recognizer will output a "rejection." What happens when an out-of-vocabulary utterance (utterance not covered by the grammar) is spoken by the user? The recognizer is still trying to match it to the set of sentences described by the grammar and will output the best match, or rejection. Since the grammar does not contain the spoken sentence, the best match of the recognizer in the out-of-vocabulary case is always erroneous; therefore, to improve the accuracy of the grammar, we need to expand it to cover as many possible sentences that the user can utter. For example, in the yes/no grammar, we incorporated many variations of "yes" and "no," such as "yeh," "sure," and "nope." Another way to control the accuracy is to improve the rejection mechanism to guarantee that out-of-vocabulary utterances will be rejected.

Based on these considerations, we deployed the following methodology in the system design described below: 1) improving rejection mechanisms for confirmation and other grammars, 2) using confirmations as the way to control the larger grammar's accuracy, and 3) using recordings to capture the unexpected and problematic spoken responses.

Improved Rejection Mechanisms for Confirmation and Other Grammars

We incorporated a garbage model in the yes/no grammar used for confirmations in our application. The garbage model was designed to match out-of-vocabulary utterances [17,18], specifically the corrections users are frequently providing instead of a negative confirmation (ie, those utterances that do not represent "yes" or "no" answers), for example

> ***System prompt:*** *Was that your left shoulder?*
>
> ***User:*** *No, right shoulder.*

(See also Multimedia Appendix 2.)

We used rejection criteria based on a combination of recognition score and garbage model scoring to control the overall accuracy of this grammar. In the case when, for a given spoken utterance, the recognizer outputs a hypothesis that is a part of a garbage model, this utterance is rejected. Also, if the hypothesis is not part of a garbage model, but has a low recognition score, it is rejected as well.

Using Confirmations as the Way to Control the Larger Grammar's Accuracy

The grammars that are substantially larger than yes/no are also those for which we can expect more ASR errors and out-of-vocabulary utterances. Those are grammars like the body part grammar or the symptoms grammar when, without substantial data collection, we cannot accurately predict all possible ways the users will answer the questions "Where does it hurt?" or "What's your most disturbing symptom?" For such grammars, we use the confirmation mechanism to control the overall accuracy of the data we capture. The result is considered captured only if the user answers "yes" to the confirmation question, reducing the error rate for the dialogue units with larger grammars to the level of yes/no grammar.

Using Recording to Capture the Unexpected and Problematic Spoken Responses

In almost all applications of ASR, especially in spontaneous speech tasks, the recognizer vocabulary cannot cover all occurring words. There is always a significant amount of out-of-vocabulary words even when the vocabulary size is very large. For example, an out-of-vocabulary input can occur when the user is trying to answer the question "Where does it hurt?" with a word or a phrase that is not covered by body part grammar. Since, by definition, out-of-vocabulary input cannot be recognized correctly, it can lead to either a rejection by the recognizer or to misrecognition to be followed (presumably) by a negative confirmation. In both cases, as will be explained in the next section, the system will detect speech recognition failure, will describe in the next prompt what the in-vocabulary choices are, and will offer the user to say a key phrase like "none of those," indicating that the choices offered do not adequately describe the user's intended input. When the user chooses the "none of those" option, the system will record the user's input and flag it as "transcription is needed" for later processing.

> ***System prompt:*** *Was that your left shoulder?*
>
> ***User:*** *No.*
>
> ***System prompt:*** *Sorry about that. Let's try it this way. Please choose carefully a body part from the following list that best describes the location of your pain, and just say it. If none of the locations match, please say "none of those." Here is the list: abdomen [pause], ankles…*
>
> ***User (interrupts):*** *None of those.*
>
> ***System prompt:*** *OK. Let me just record your answer. Please describe the location of your pain in your own words.*
>
> ***User:*** *[description of pain location]*
>
> ***System prompt (after recording is finished):*** *Thanks, I got that. Let's move on.*

(See also Multimedia Appendix 3.)

The same mechanism of fall-back to recording instead of recognition is used after several repeated recognition failures,

like in a case when the user does not chose the "none-of-those" option, while repeating the out-of-vocabulary input.

In the following example, the user is trying to answer the question about type of pain with an unexpected out-of-vocabulary input.

> ***System prompt (following negative confirmation):*** *Sorry about that. Let's try one more time. Please choose carefully a word from a following list that best describes your pain and just say it. If nothing matches, please say "none of those." Here is the list: aching [pause], burning [pause], dull…*
>
> ***User (interrupts with out-of-vocabulary input):*** *I have high fever.*
>
> ***System prompt:*** *Was that "deep?"*
>
> ***User*** *: No.*
>
> ***System prompt:*** *Sorry, I am having trouble. Let me just record your answer. Please describe how your pain feels in your own words.*
>
> ***User:*** *I am burning with fever.*
>
> ***System prompt (after recording is finished):*** *Thanks, I got that. Let's move on [pause]. Is the pain constant…?*

(See also Multimedia Appendix 4.)

Flexible Level of User Support

The flexible level of user support that is intended to satisfy both the novice and the experienced user is achieved by prompt design, context-sensitive help, detecting speech recognition failures, and dialogue personalization, described below.

Prompt Design

The system prompts are designed to provide an appropriate level of support to the user. For example, the initial prompt for the pain location dialogue unit is "Where does it hurt?" [pause]. For example, your head, stomach, or back? [pause]. Remember, if youdon't know how to answer this question, just say 'I need help.'" The pauses in this prompt are designed to encourage the experienced user to interrupt the prompt with the answer (most experienced users interrupt after the initial prompt), while providing more information (in this case, examples of possible answers) for the inexperienced user who hesitates to answer immediately. It also reminds the user to ask for help if it is still not clear what can be said as an answer. Multimedia Appendices 5, 6, and 7 contain the recordings illustrating the different user experiences with this prompt.

Context-Sensitive Help

Although participants may have received some training at their orientation session, it is unreasonable to expect them to retain this information for the whole duration of the trial, which can last for months. Therefore, for every question in the Pain Monitoring Voice Diary help information is provided upon the user's request, describing and clarifying the current question, and, in some cases, enumerating the possible answers the caller can choose from while, in other cases, giving more examples of possible answers. For example, if the caller asks for help after the "Where does it hurt?" question, the system will provide a very elaborate help prompt that lists different body parts the user can say (pausing shortly after each one to encourage the user to interrupt if the user knows what to say). It also reminds the user about the "none of those" option:

Okay. Here is the help information. At this point I need to find out the part of your body that hurts the most. Please choose carefully a body part from the following list that best describes the location of your pain, and just say it. If none of them matches, please say "none of those." Here is the list: abdomen [pause], ankles [pause], back [pause]...toes [pause]. Which one is it?

The information provided during these explicit requests for help closely follows the information the user received during the enrollment process. The recording in Multimedia Appendix 8 illustrates a case of help request.

Detecting Speech Recognition Failures

Even when the user has not asked for help explicitly, the dialogue is designed to detect the user's repeated failures and provide more support. When the system experiences recognition problems, such as rejection or silence, it will re-prompt the user for the same question. The re-prompts are designed as an escalating list, providing increasingly more information and progressively constraining the user as more such errors are detected. For example, if the user's utterance is rejected by the recognizer after the initial prompt "Where does it hurt? [pause]. For example, your head, stomach, or back? [pause]. Remember, if you don't know how to answer this question, just say 'I need help,'" the system will re-prompt for the same information with "I didn't get that. Let's try it this way. Please tell me the part of your body that hurts the most. Remember, you could always say 'I need help.'" The second prompt skips the pauses and reminds the user to ask for help if needed, and it also clarifies the question ("body part that hurts the most").

Another case when the system detects that something went wrong with speech recognition is when the user says "no" to a confirmation question:

> ***System prompt:*** *Was that your left shoulder?*
>
> ***User:*** *No.*
>
> ***System prompt:*** *Sorry about that. Let's try it this way. Please choose carefully a body part from the following list that best describes the location of your pain, and just say it. If none of them matches, please say "none of those." Here is the list: abdomen [pause]…. Which one is it?*

(See also Multimedia Appendix 9.)

Since the user disconfirmed the recognized body part, the system detects a recognition problem and gives the user more information on how this question can be answered to minimize the out-of-vocabulary utterance rate.

Dialogue Personalization

Data capture is a unique dialogue application since the users not only call the system many times during the trial, but they also identify themselves at the beginning of each session. This provides a system with an opportunity to personalize both the

content of the current session (what are the data to be collected) as well as the style (how to ask for these data) based on the results of the previous sessions. As shown in Figure 2, in our system we took advantage of a larger inter-session context by designing two types of data collection sessions: normal and follow-up. The follow-up session type is deployed if the subject reported a high level of pain in the previous session. The follow-up session differs from the normal one not only by the additional questions the patient is asked, such as if and when the subject took the medication, but also by the format of the questions. If, in the previous session, the subject reported pain in the left shoulder, in the follow-up session the question will be "Is the pain still in your left shoulder?" (see Multimedia Appendix 10). This format of "reminding" prompts was used for pain location and pain type dialogue units, and it was designed to possibly shorten the dialogues and also provide the user comfort and a feeling of continuity in using the system. Textbox 2 presents a transcription of a follow-up session (it is the follow-up for the normal session shown in Textbox 1) and Multimedia Appendix 11 contains its recording. Table 2 describes an automatically generated report for this follow-up session.

Textbox 2. Transcription of a follow-up dialogue

System: Your PIN, please.

Caller: ****

System: Hi, Esther, This is your follow-up call…. Are you still having pain?

Caller: Yeh.

System: Did you take your pain medication today?

Caller: Yes.

System: What time did you take the medication?

Caller: Two hours ago.

System: Was that around 9:45 am?

Caller: Yeh.

System: Did the medicine diminish your pain?... Please…

Caller: Yeh.

System: Please rate your pain on a scale of 0 to 10, with 0 being no pain and 10 being the worst possible pain.

Caller: Four.

System: Was that 4?

Caller: Yeah.

System: Is the pain still in your left shoulder?

Caller: Yes.

System: Does it still feel sharp?

Caller: No.

System: How does the pain feel?

Caller: It's dull.

System: Was that "dull?"

Caller: Yeah.

System: Is the pain constant or does it come and go?

Caller: It's constant.

System: Do you have any other symptoms?

Caller: No.

System: Would you like to leave a message for the staff?

Caller: No.

System: Thanks for your cooperation. I've got all the information that I need. I'll call you again tomorrow. Goodbye.

Table 2. Automatically generated report for the follow-up session in Textbox 2

	Captured Value	Confirmed (yes/no)	Confidence Score
PIN	*****	no	74
Are you in pain?	yes	no	85
Medication taken?	yes	no	76
Medication time	9:45 am	yes	69
Medication helped?	yes	no	75
Pain rating	4	yes	87
Pain location	left shoulder	yes	87
Pain type	dull	yes	86

Feasibility Study

Experimental evaluation of usability of the Pain Monitoring Voice Diary was performed with 24 volunteers (8 females, 16 males), mostly City College students. The volunteers were instructed either to refer to their past injury/sickness/pain episode experiences or to choose a scenario out of a set of nine that included scenarios for migraine pain, back pain, post-surgical pain (knee injury), arthritis and others. The goal of this evaluation was to prove the feasibility of data capture through dialogue and validate the assumptions underlying dialogue design.

The volunteers were asked to contribute 10 sessions with the system over a period of 2 weeks; in practice, the number of sessions per subject ranged from 1 to 20. There was no formal training session provided; instead, once enrolled (through a website) the subjects received an email notification with their PIN and general information about the system. The subjects were asked to either relate to pain episodes in their past while answering the system's questions or use as a guidance one of nine provided medical scenarios compiled by a pain specialist, ranging from migraines and back pain to post-surgery pain (knee injury) and cancer- and chemotherapy-related afflictions.

Results

Dialogue Evaluation

From 24 volunteers we collected a total of 177 dialogue sessions: 171 sessions were completed, while the caller hung up in 6 sessions; 66 of the completed sessions were the follow-up type. There were a total of 2437 dialogue turns, where dialogue turn corresponds to one system prompt and one user utterance. The data capture rate, measuring the percentage of slots filled automatically, was 98%, while the other 2% were flagged for transcription. Data capture rate is not a direct measure of ASR accuracy since slots are not necessarily filled after the first attempt. Among the utterances sent to transcription, where the user had opted for the "none of those" option, 70% corresponded to the type of pain slot, 20% to the symptoms slot, and 10% to the body part slot, indicating that those are the grammars with the highest out-of-vocabulary rate. Since all captured slots were confirmed, to evaluate the accuracy of the captured data we had to evaluate the accuracy of the confirmation grammar used to recognize confirmation utterances. Among 859 confirmation utterances, 10 were misrecognized, leading to a 98.8% data accuracy rate. The rejection parameters in the grammar were tuned for an equal misrecognition/rejection rate and, indeed, the number of rejections among the 859 confirmation utterances was 11.

Table 3. Average dialogue session statistics (figures in parentheses are standard deviations)

Session duration (s)	99.34 (45.92)
Number of dialogue units per session	7.65 (2.48)
Duration of dialogue unit (s)	12.99 (2.7)
Dialogue turns per dialogue unit	1.86 (0.43)
Percentage of task-oriented turns	82% (15.4)
Percentage of interrupted prompts	68% (13)
Time duration of a dialogue turn (s)	6.97 (1.3)
Time duration of a dialogue turn when interruption was disabled	10.63 (1.5)

Table 3 shows other metrics [19] derived from dialogues. The high standard deviations of session duration and dialogue units per session are due to the extensive variability of dialogue sessions. The sessions not only differ by type (normal and follow-up), but there is also branching within the same type application (eg, some of the users report symptoms, while others don't, some take medications). In addition, there is a great variability due to ASR errors and different possibilities inherent in the design of the call flow (eg, caller initiated help requests, speech recognition error handling such as re-prompts, negative confirmations).

The high standard deviations in caller utterances per dialogue unit and dialogue unit duration are due to the fact that not all dialogue units are created equal. For example, the “Are you in pain?” dialogue unit can fill a slot with a single “yes”/“no” utterance, while the pain location unit requires at least two dialogue utterances (body part and confirmation) if speech recognition does not fail, and more if it does.

Percentage of task-oriented dialogue turns (82%) (those dialogue turns that are *not* due to speech recognition errors or caller help requests) is a measure of dialogue efficiency: if there were no errors and help requests at all, it would be 100%. The prompts in the dialogue were designed to be interrupted by experienced callers. To quantify the use of interruption, we computed the percentage of interrupted prompts (68%). To quantify how far in the prompts the interruption occurs, we computed the average duration of dialogue turn (6.97 s), and compared it to the reference of average duration of dialogue turn (10.63 s) when the interruption feature was disabled.

Evaluation of Flexible Level of User Support

One of the goals of the dialogue design described above was to have a flexible and adaptive user support for different types of users, providing short prompts and efficient call flow for experienced users and more detailed information in a troublesome situation for novice users. To evaluate the efficiency of the dialogues as a function of user proficiency, we divided the sessions into seven classes according to the sequential order of the session with the same user. Table 4 shows some statistics of the classes. For example, class A contains all the first sessions each of the 24 users had, with a total of 308 dialogue turns; class G contains all the sessions (whose ordinal number was 10 and above) for which the users had at least nine sessions previously completed.

Table 4. Dialogue sessions divided according to the call order

Class Name	Call Order	Sessions	Turns
A	1	24	308
B	2	19	302
C	3	15	206
D	4, 5	27	380
E	6, 7	23	324
F	8, 9	21	305
G	10+	43	612

Figures 3, 4, and 5 illustrate the average dialogue turn duration, average percentage of interrupted prompts, and average percentage of task-oriented prompts, respectively, for the classes outlined in Table 4. The differences between the seven session classes for the three metrics shown are statistically significant as tested by ANOVA [20] ($F = 49.33$, 50.40, and 50.40, respectively; $df = 6$; $P < .001$). The results in Figures 3 and 4 confirm the assumptions of the dialogue design: the prompts were designed to be interrupted by experienced users, and, indeed, the results indicate that the more experienced the user is, the more often and earlier she will interrupt. The novice user only interrupts in 59% of the prompts, with an average turn duration of 7.7 s, while users that had more than nine sessions completed interrupted in 73% of the prompts, with an average dialogue turn duration of 6.5 s. Figure 5 shows that with experience the users become more efficient with the system, as measured by the percentage of task-oriented dialogue turns: for novice users this percentage averages 75%, increasing to around 81% after just one previous session was completed, and up to 86% after at least nine sessions were previously completed.

Figure 3. Average turn duration, in seconds, for dialogues in classes A to G (a dialogue turn corresponds to one system prompt and one user utterance; classes indicate increased experience, with A being the most inexperienced users; error bars indicate 95% confidence intervals)

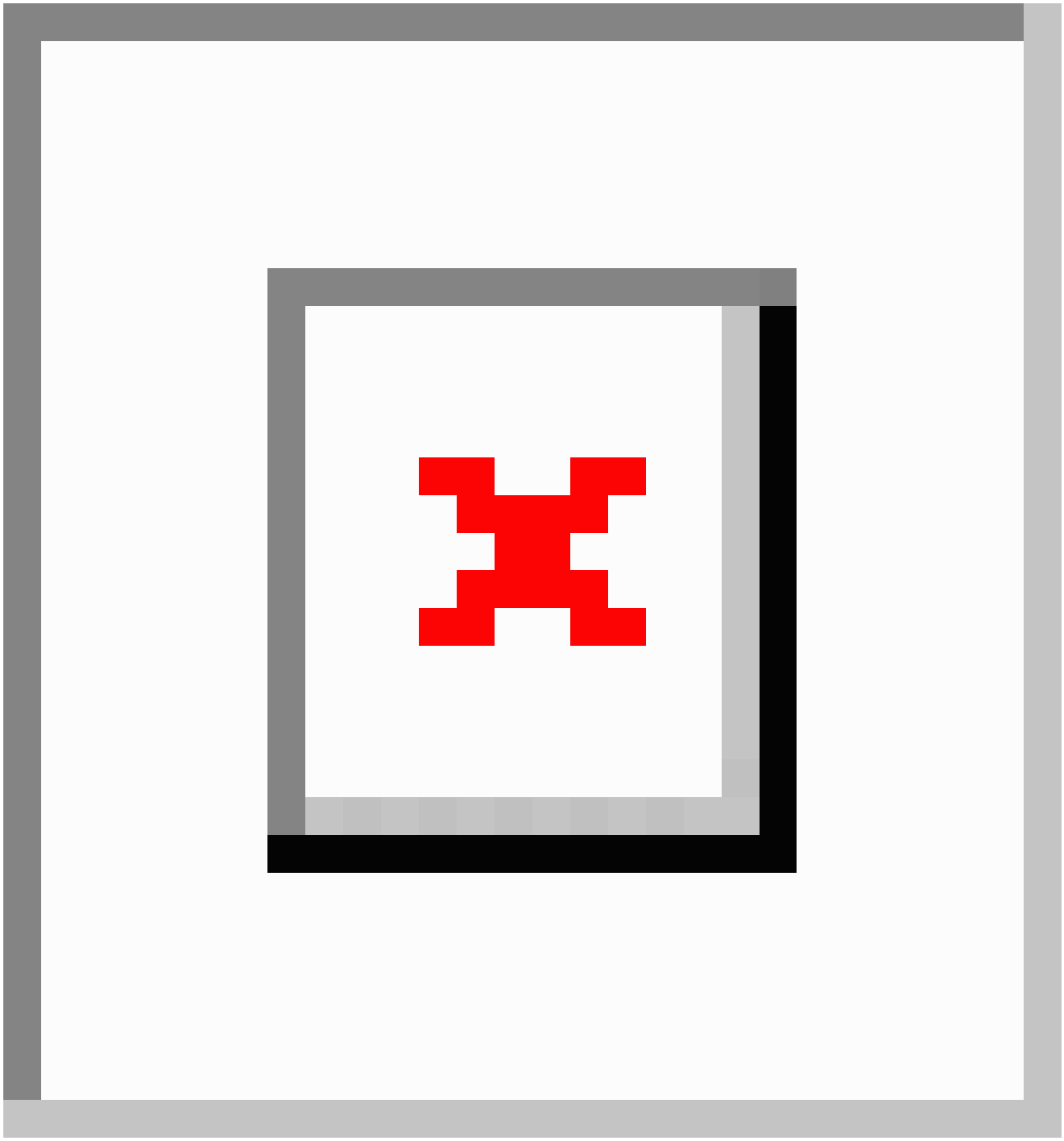


Figure 4. Percentage of interrupted prompts for dialogues in classes A to G (error bars indicate 95% confidence intervals)

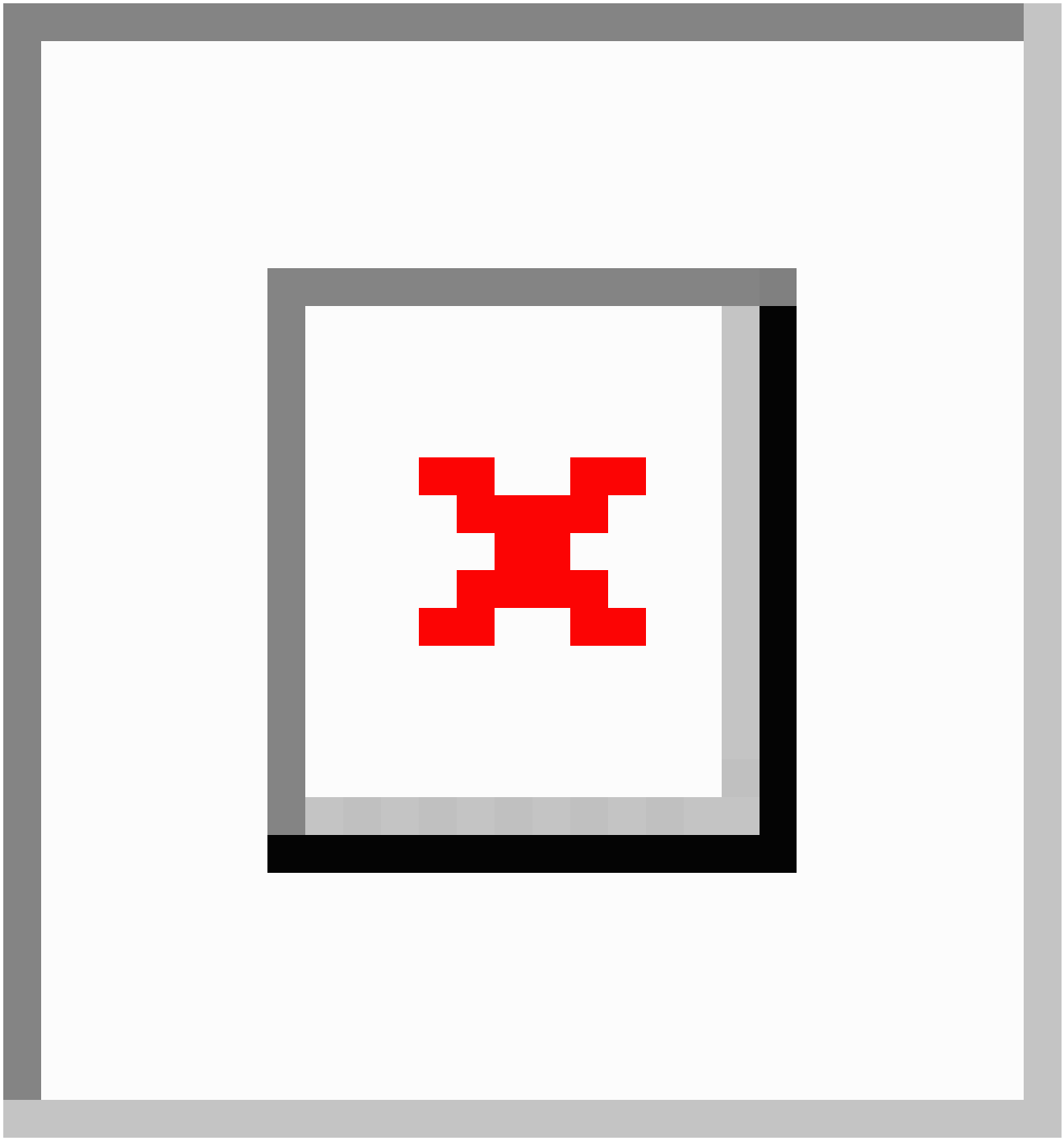
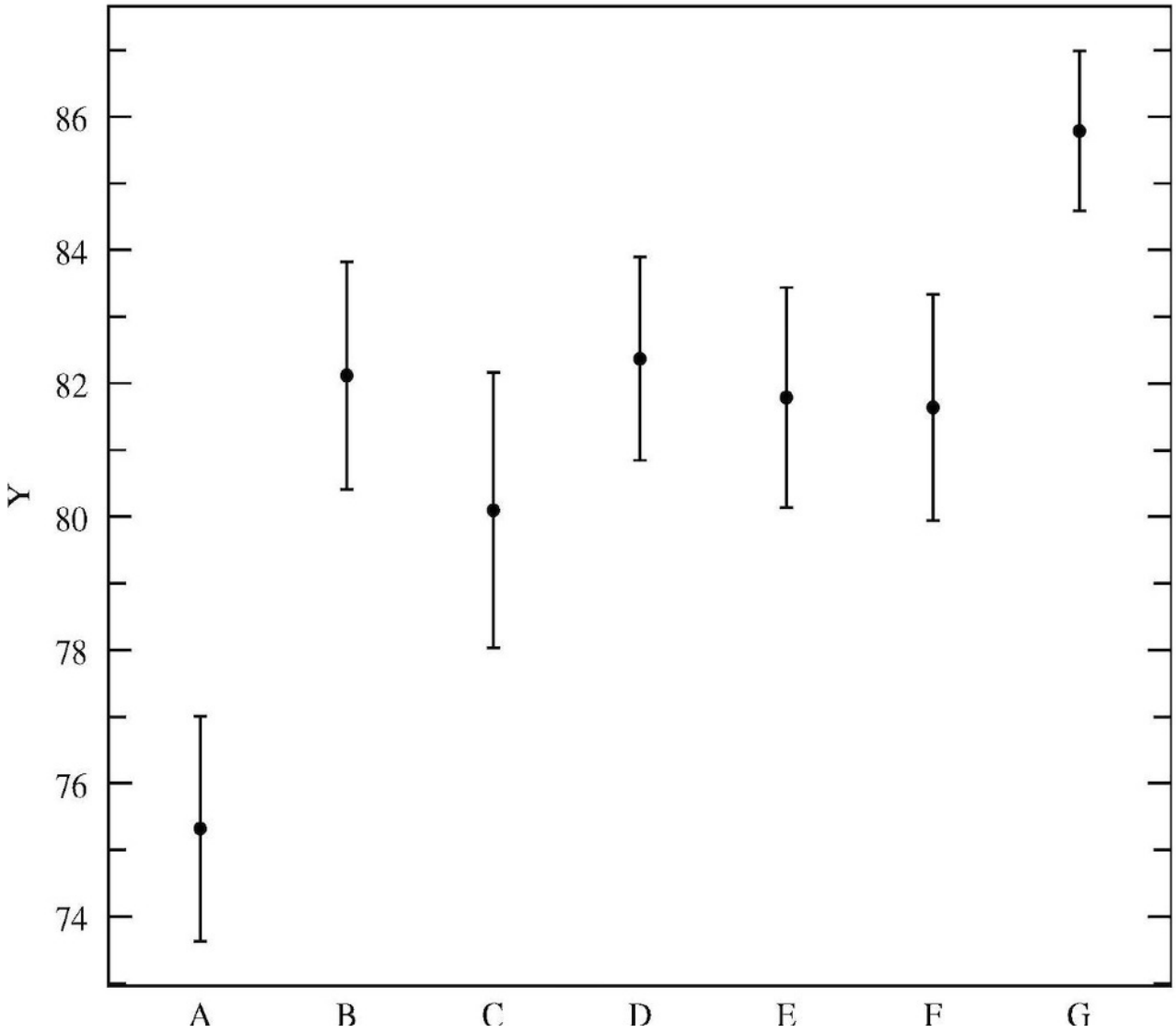


Figure 5. Percentage of task-oriented turns for dialogues in classes A to G (error bars indicate 95% confidence intervals)



Discussion

In this paper we present a new methodology for real-time data collection of information from patients for health, behavioral, and lifestyle studies and monitoring.The voice data collection system is intended to facilitate real-time collection of information from patients via automated speech telephony delivery. A flexible self-report system gives patients the freedom to choose to use a phone as a device that meets their preferences, schedules, or limitations. The system monitors and tracks data collection compliance and generates real-time notifications and alerts for participants and administrators. The system provides on-demand/online reports to enhance informed decision making to improve patient care. The reports can be tailored to profile patient function over time and highlight clinically meaningful changes in health status.

Some of the major challenges in designing such a system include maintaining adequately high accuracy of the captured data while guaranteeing a satisfactory user experience. We described dialogue design for a Pain Monitoring Diary that targets these challenges. In particular, to control the accuracy of automatically captured data, we tuned the parameters of the garbage model–based confirmation grammar to reliably reject out-of vocabulary utterances; every captured value was explicitly confirmed using this high-accuracy confirmation grammar, and, if needed, hard to recognize or out-of-vocabulary answers were recorded and flagged for later transcription. To provide a flexible level of user support, we used a variety of methods, including designing the prompts to be interrupted by experienced users while carrying enough information for novices; providing the user with on-demand, context-dependent help; and detecting troublesome situations and guiding the user through them with more informative prompts or resorting to recording of the answer provided and flagging it for later transcription. The results of the feasibility study indicate that desired accuracy of data can be achieved with still a high degree of automation (98.8% data accuracy with 98% automation). The users were capable of using the flexible interface, with the sessions becoming more and more efficient as the users' experience increased, both in terms of session duration and avoidance of troublesome dialogue situations. The adaptation data shown in Figures 3 to 5 suggest that even some rudimentary experience with the system increases the session efficiency significantly. While the subjects

in the feasibility study did not receive any training or orientation session with the system, these results suggest that there could be significant value in conducting such sessions, which may provide users with the experience to jump start in the middle.

Finally, we would like to stress that one of the major weaknesses of the speech system is its single modality. Protocols that rely heavily on graphics and visual formats that cannot be completely replicated in speech are probably a poor match for this technology. Therefore, we do not see the spoken dialogue–based data collection as a replacement for existing data collection methodologies, but only as another choice for health care providers and researchers.

Acknowledgments

The Pain Monitoring Voice Diary (PMVD) system was developed by Spacegate, Inc under the brand name SpeechMatrix, which is currently scheduled for validation trials with Beth Israel, New York Cancer Center. The project described was supported by the grant “Automated Speech Real-Time Patient Data Collection” from National Institutes of Health / National Cancer Institute. Its contents are solely the responsibility of the authors and do not necessarily represent the official views of the NIH.

Conflicts of Interest

None declared.

Multimedia Appendix 1

Recording of a dialogue session with the system in normal mode (Transcription of this session appears in Textbox 1.) [WAV file (Waveform Audio Format), 1.6 MB - jmir_v8i4e30_app1.wav]

Multimedia Appendix 2

Recording of several dialogue exchanges with the system illustrating the rejection of out-of-vocabulary utterances [WAV file (Waveform Audio Format), 264 KB - jmir_v8i4e30_app2.wav]

Multimedia Appendix 3 and 4

Recordings of several dialogue exchanges illustrating the use of recording to capture out-of-vocabulary or problematic user inputs [WAV file (Waveform Audio Format), 832 KB - jmir_v8i4e30_app3.wav] [WAV file (Waveform Audio Format), 676 KB - jmir_v8i4e30_app4.wav]

Multimedia Appendix 5-7

Recordings of different user experiences with the “Where does it hurt?” prompt [WAV file (Waveform Audio Format), 96 KB - jmir_v8i4e30_app5.wav] [WAV file (Waveform Audio Format), 156 KB - jmir_v8i4e30_app6.wav] [WAV file (Waveform Audio Format), 296 KB - jmir_v8i4e30_app7.wav]

Multimedia Appendix 8

Recording illustrating system behavior when help is requested [WAV file (Waveform Audio Format), 968 KB - jmir_v8i4e30_app8.wav]

Multimedia Appendix 9

Recording illustrating system behavior when it detects a problematic situation [WAV file (Waveform Audio Format), 660 KB - jmir_v8i4e30_app9.wav]

Multimedia Appendix 10

Recording of reminder style prompts in follow-up mode [WAV file (Waveform Audio Format), 136 KB - jmir_v8i4e30_app10.wav]

Multimedia Appendix 11

Recording of a dialogue session with the system in follow-up mode (Transcription of this session appears in Textbox 2; this session is the follow-up for the normal session in Textbox 1 and recorded in Multimedia Appendix 1.) [WAV file (Waveform Audio Format), 1.4 MB - jmir_v8i4e30_app11.wav]

References

1. Collins RL, Morsheimer ET, Shiffman S, Paty JA, Gnys M, Papandonatos GD. Ecological momentary assessment in a behavioral drinking moderation training program .
2. O'Connell KA, Gerkovich MM, Cook MR, Shiffman S, Hickcox M, Kakolewski KE. Coping in real time: using Ecological Momentary Assessment techniques to assess coping with the urge to smoke. Res Nurs Health 1998 Dec;21(6):487-497. [Medline: 99053529]
3. Hufford MR, Shifford S. Capturing real-time, real-world quality of life data using ecological momentary assessment. Quality of Life Newsletter 2000 Jul;25:1-3.
4. Turk DC, Melzack R, editors. Handbook of pain assessment. 2nd edition. New York: The Guilford Press; 2001.
5. Smyth J, Stone AA. Ecological momentary assessment research in behavioral medicine. Journal of Happiness Studies 2003;4(1):35-52. [doi: 10.1023/A:1023657221954]
6. Smyth J, Wonderlich S, Crosby R, Miltenberger R, Mitchell J, Rorty M. The use of ecological momentary assessment approaches in eating disorder research. Int J Eat Disord 2001 Jul;30(1):83-95. [Medline: 21332654] [doi: 10.1002/eat.1057]
7. Stone AA, Shiffman S, Schwartz JE, Broderick JE, Hufford MR. Patient non-compliance with paper diaries. BMJ 2002 May 18;324(7347):1193-1194. [Medline: 22011654] [doi: 10.1136/bmj.324.7347.1193]
8. Paul J, Seib R, Prescott T. The Internet and clinical trials: background, online resources, examples and issues. J Med Internet Res 2005;7(1):e5 [FREE Full text] [Medline: 15829477] [doi: 10.2196/jmir.7.1.e5]
9. Collins RL, Kashdan TB, Gollnisch G. The feasibility of using cellular phones to collect ecological momentary assessment data: application to alcohol consumption. Exp Clin Psychopharmacol 2003 Feb;11(1):73-78. [Medline: 22509271] [doi: 10.1037/1064-1297.11.1.73]
10. Black L, McTear M, Black N, Harper R, Lemon M. The voice-logbook: integrating human factors for chronic care system. Presented at: ICSLP; October, 2004; Jeju Island.
11. Black L, McTear M, Black N, Harper R, Lemon M. Evaluating the DI@L-log system on a cohort of elderly, diabetic patients: results from a preliminary study. Presented at: Interspeech; 2005; Lisbon, Portugal.
12. Levin E, Levin A. Spoken dialog system for real-time data capture. Presented at: Interspeech; 2005; Lisbon, Portugal.
13. Levin E, Levin A. Dialog design for user adaptation. Presented at: ICASSP International Conference on Acoustics; May 14-19, 2006; Toulouse, France.
14. Davis MP, Walsh D. Cancer pain: how to measure the fifth vital sign. Cleve Clin J Med 2004 Aug;71(8):625-632. [Medline: 23564451]
15. Daut RL, Cleeland CS, Flanery RC. Development of the Wisconsin Brief Pain Questionnaire to assess pain in cancer and other diseases. Pain 1983 Oct;17(2):197-210. [Medline: 84069392] [doi: 10.1016/0304-3959(83)90143-4]
16. McTear M. Spoken dialogue technology – towards the conversational user interface. London: Springer Verlag; 2004.
17. Rose R, Paul D. A hidden Markov model based keyword recognition system. Presented at: Conference on Acoustics, Speech and Signal Processing; 1990; Albuquerque, New Mexico.
18. Manos A, Zue V. A segment-based spotter using phonetic filler models. Presented at: ICASSP International; April 1997; Munich, Germany.
19. Walker MA, Litman DJ, Kamm CA, Abella A. PARADISE: a framework for evaluating spoken dialogue agents. Proceedings of ACL/EACL 35th Annual Meeting of the Association for Computational Linguistics. San Francisco: Morgan Kaufmann; 1997:271-280.
20. Edgington ES. Randomization tests. New York: Marcel Dekker Inc; 1986.

Abbreviations

ASR: automated speech recognition
EMA: ecological momentary assessment
IVR: interactive voice response
PDA: personal digital assistant

submitted 28.08.06; peer-reviewed by W Slack; comments to author 03.11.06; revised version received 19.11.06; accepted 24.11.06; published 11.12.06.

Please cite as:
Levin E, Levin A
Evaluation of Spoken Dialogue Technology for Real-Time Health Data Collection
J Med Internet Res 2006;8(4):e30
URL: http://www.jmir.org/2006/4/e30/
doi:10.2196/jmir.8.4.e30
PMID:

Original Paper

Does the Quality of the Working Alliance Predict Treatment Outcome in Online Psychotherapy for Traumatized Patients?

Christine Knaevelsrud[1], PhD; Andreas Maercker[2], MD, PhD

[1]Treatment Centre for Torture Victims, Berlin, Germany

[2]Institute of Psychology, University of Zurich, Zurich, Switzerland

Corresponding Author:
Christine Knaevelsrud, PhD
Treatment Centre for Torture Victims
GSZ Moabit
Turmstr. 21
10559 Berlin
Germany
Phone: +0049 177 7894682
Fax: +0049 303 0614371
Email: c.knaevelsrud@bzfo.de

Abstract

Background: The provision of online counseling and online therapy is steadily increasing. The results of a number of controlled trials investigating the efficacy of online approaches indicate that some of these new treatment alternatives might indeed be effective. Yet, little is known about how the therapeutic relationship (or working alliance) evolves over the Internet and whether it influences treatment outcome as it does in traditional face-to-face therapy. The working alliance has been defined as the extent to which a patient and a therapist work collaboratively and purposefully and connect emotionally.

Objective: The aim of the study was to investigate the quality and predictive relevance of the therapeutic alliance for patients receiving a short-term, Internet-based, cognitive-behavioral therapy program for posttraumatic stress reactions.

Methods: After rigorous screening for exclusion criteria of high dissociative tendencies, risk of psychosis, and suicidal tendencies, 48 patients, who had experienced a traumatic event in the past, were included in the online treatment study. The short form of the Working Alliance Inventory (WAI-S) was administered at the fourth treatment session. The relevance of the therapeutic relationship for treatment outcome was assessed in terms of residual gain from pretreatment assessment to the end of treatment. The revised Impact of Event Scale (IES-R) and the depression and anxiety subscales of the Brief Symptom Inventory (BSI) were used to assess treatment outcome.

Results: A total of 48 participants were included in the analysis. Overall, high alliance scores were found. In contrast to previous studies of conventional face-to-face therapy, there was only a low to modest association (.13 to .33) between the quality of the therapeutic relationship and treatment outcome.

Conclusion: High alliance scores indicate that it was possible to establish a stable and positive therapeutic relationship online. However, the therapeutic relationship was found to be a less relevant predictor of the therapy outcome than in face-to-face approaches. We discuss whether this finding can be attributed to methodological reasons such as the restricted range of alliance ratings obtained or the time of administration of the WAI-S, or whether the therapeutic relationship might be less relevant to the treatment outcome of online therapy approaches.

(J Med Internet Res 2006;8(4):e31) doi:10.2196/jmir.8.4.e31

KEYWORDS

Online therapy; Internet; working alliance; therapeutic relationship; psychotherapy; psychotherapeutic processes; professional-patient relations; treatment outcome

Introduction

Recent developments in communication technology have opened up new therapeutic possibilities that challenge our understanding of psychotherapy. While the academic debate continues as to whether online treatments might present an acceptable alternative to face-to-face therapy, real life has already decided. “Researchers can no longer discuss online counseling as an

intervention method that will take shape in the future—the future is now" [1]. Internet-based treatment approaches have already been developed for a wide range of clinical disorders including depression, eating disorders, anxiety disorders, and substance abuse, as have interventions targeting relationship problems, adjustment disorders, and work-related burnout, and the numbers are expected to increase [2]. The numbers of empirical studies investigating the efficacy of online approaches are growing apace, and results indicate that some of these new treatment alternatives might indeed be effective (see [3] for a review). However, an important question remains largely unanswered: What contributes to therapeutic change? To date, virtually no studies have focused on the processes underlying online therapy [4]. Thus, it is not clear whether online therapy is based on factors and mechanisms similar to those that are responsible for therapeutic change in face-to-face therapy or whether we need to redefine our understanding of the underlying processes when considering online therapy.

The Therapeutic Relationship

The quality of the therapeutic relationship, or working alliance, has been demonstrated to be especially important in predicting the outcome of psychotherapy. The working alliance has been defined as the extent to which a patient and a therapist work collaboratively and purposefully and connect emotionally [5]. Research reviews have consistently reported a positive relationship across studies between the quality of the therapeutic alliance and therapy outcome, although there are some instances where the working alliance fails to predict outcome or where associations are nonsignificant [5-10]. In their meta-analysis, Martin et al [10] reported that the quality of the therapeutic alliance accounted for 22% of the variance in the rate of therapeutic success. Moreover, research has indicated that the relationship between therapeutic alliance and treatment outcome holds across several types of treatment, including cognitive-behavioral therapy (CBT) [11], interpersonal therapy [9], and psychodynamic therapy [7,11], and does not differ significantly within treatment approaches [8,9].

Online Relationships

The beneficial effects and clinical relevance of a positive working alliance have been well documented in face-to-face therapies, but almost nothing is known about how the therapeutic relationship operates online. Online therapy challenges our basic assumptions about what is needed to establish a therapeutic contact, such as (1) sharing the same physical space, (2) talking, and (3) synchronous real-time interaction [12], and it is still uncertain if online therapy provides conditions that are sufficient to establish a stable therapeutic alliance at all. Since one of the major criticisms of online therapy concerns the ambiguous nature of the therapeutic relationship, research in this field is needed. Most previous studies have focused on relational behavior in everyday online contact, with inconsistent results. These findings prompted an academic discussion between proponents of two contrasting views of the online relationship. On the one hand, Slouka [13] states that online relationships are shallow, impersonal, and unreal. Indeed, Kraut et al [14] have demonstrated that online relationships heighten depression and loneliness rather than provide fulfilling relationships. Mallen et al [1] compared Internet-based and face-to-face conversations in a randomized study and found that participants who communicated online felt less satisfied with their contact and experienced a lower degree of self-disclosure and closeness with their partner than participants in the face-to-face group. On the other hand, various other authors have shown that online contacts are just as real and intense as face-to-face relationships and that differences between online relationships and face-to-face relationships diminish over time [15]. Whitty and Gavin [16] found that the absence of social clues enhanced and encouraged the development of relationships. This is in line with prior research indicating that visual anonymity contributes to higher levels of self-disclosure and openness [17,18].

It should be noted, however, that online therapeutic contact differs markedly from arbitrary, anonymous online contact, the most important difference probably being the identity of the therapist. In online therapeutic contact, the address, telephone number, and credentials of both parties are accessible. Furthermore, the frequency of contact is predefined and there are set time limits for response. Thus, aspects such as uncertainty about the identity and honesty of the other party, which might be detrimental to establishing a trustful contact, are much reduced in online therapeutic relationships compared with anonymous online contacts.

Focusing on the working alliance online, Cook and Doyle [19] evaluated differences in client ratings of the working alliance between a small sample (N = 15) of online therapy clients and normative data from a comparable face-to-face counseling sample. They found comparable (and relatively high) evaluations of the working alliance in the online sample using the frequently applied Working Alliance Inventory (WAI) [20].

Lange et al [21] conducted an Internet-based treatment study of work-related burnout. After completing the course of treatment, patients where asked to rate the contact with their therapists: 75% of the 115 participants described the contact as personal and 88%, as pleasant; 80% rated being treated exclusively via the Internet as positive, and 70% indicated that they did not miss face-to-face contact. Cohen and Kerr [22] compared the impact of one session of face-to-face counseling with online counseling (chat) in terms of posttreatment anxiety and attitudes toward counseling. Participants (N = 24) were randomly assigned to one of the two experimental groups. Clients in both groups experienced a uniform decrease in anxiety and rated their counselors equally on expertness, attractiveness, and trustworthiness, regardless of the mode of delivery.

While data from the aforementioned studies provide valuable information and preliminary evidence that a positive working alliance can be developed through the Internet, empirical data derived from systematic exploration of the online therapeutic relationship remain sparse. Thus, it is essential to investigate whether it is possible to develop a therapeutic alliance in the absence of visual and auditory cues and whether the working alliance has the same predictive value in online treatment as in face-to-face therapy.

Research Questions

The present study aims to replicate prior findings concerning the relationship between the working alliance and treatment outcome in face-to-face therapy. It was hypothesized that the baseline psychopathology would be inversely associated with the patients' assessment of the therapeutic alliance. Furthermore, it was hypothesized that the quality of the online therapeutic alliance would predict the residual gain from pretreatment assessment to end of treatment. We expected the patients' ratings of the alliance to be more highly correlated with therapy outcome than the therapists' ratings, and the patients' and therapists' assessments of the therapeutic alliance to be only moderately related. Overall, we expected that it would be possible to establish a positive and stable therapeutic relationship online, characterized by high scores on the WAI. The present study is part of a larger study with random assignment to a treatment group or a waiting-list control group [Knaevelsrud and Maercker, in preparation]. Based on the research questions chosen, only the data from the treatment group were used in the following analyses.

Methods

Recruitment

Participants were recruited by means of radio and newspaper advertisements as well as advertisements posted on websites for different groups (eg, crime victims, sexual abuse victims, bereaved parents). To be included in the study, participants had to (1) have experienced a traumatic event that occurred at least one month prior to treatment and that met the criteria specified in the DSM-IV [24], (2) be 18 years or older, (3) not exceed the cutoff scores for dissociation and psychosis (see exclusion criteria), (4) not abuse alcohol or other drugs, (5) not consume neuroleptics, (6) be fluent in written German, and (7) not be receiving treatment elsewhere.

A total of 498 potential participants showed interest in the treatment; 68% (N = 337) returned the screening questionnaires; 73% (N = 246) were excluded based on the exclusion criteria. In total, 91 patients participated in this study (48 in treatment group; 43 in control group).

Potential patients browsed through the Interapy website, which provided information about (1) posttraumatic stress reactions, (2) the study and its inclusion criteria, (3) the treatment, (4) the therapists and supervisors, and (5) other treatment alternatives. Potential participants were sent screening questionnaires by email. Those who passed the screening received an informed consent document. Participants were required to sign and return this document, indicating that they had been informed about the aim and procedures of the research project and were willing to take part in it. Based on a computer-generated randomization list, they were randomly assigned to the waitlist-control group or treatment group. Patients who were excluded from the study were provided with information on where they could receive treatment elsewhere.

To gather miscellaneous information, including the time since the trauma, education level, degree of computer and Internet experience, and typing skills, a short checklist was administered.

Therapists

Two therapists conducted the treatment. Both were female psychologists who had received special training in the application of writing assignments for the treatment of posttraumatic stress disorder (PTSD). One was trained in cognitive-behavioral psychotherapy. Their average age was 33 years. The therapists participated in weekly supervision sessions.

Exclusion Criteria

Dissociation

Dissociative symptoms were tapped using the Somatoform Dissociation Questionnaire (SDQ-5) [25]. The scale consists of five items, which are rated on a 5-point Likert scale (1 = not at all, 5 = very often). The internal consistency of the SDQ-5 is good (α = .80). Participants who scored above the cutoff score were excluded from the treatment.

Risk of Psychosis

Risk of psychosis was measured using the Dutch Screening Device for Psychotic Disorder [26]. This seven-item inventory has high internal consistency (α = .82) and is a good predictor of psychotic episodes. In a Dutch study, a high level of agreement was found between the self-reports of 33 patients and their clinicians' reports on them (α = .85). Since no German norm group exists as yet, the data from the Dutch norm group were used. Participants were excluded if they scored above the cutoff score. Participants were also excluded if they indicated the use of neuroleptics.

Risk of Suicide

Suicidal intentions and risk of suicide were measured using the Suicide Risk Assessment (SRA) [27], a six-item, self-report questionnaire designed to capture suicidal tendencies. It consists of questions tapping suicidal plans, previous suicide attempts, and current suicidal intentions.

Treatment

Patients were sent two weekly 45-minute writing assignments over a five-week period (10 essays in total). Before and after the treatment, participants completed a set of questionnaires measuring the treatment effect. The therapy consisted of three treatment phases: (1) self-confrontation, (2) cognitive reconstruction, and (3) social sharing. After the fourth writing session, which constituted the end of the first treatment phase, the short form of the WAI (WAI-S) was administered. The treatment procedure is described in detail by Lange et al [23] and will only be outlined in brief here.

First Phase: Self-Confrontation

At the beginning of the treatment, participants received psychoeducation on the mechanisms of exposure. In the first phase, the therapists helped the patients to focus on the most painful images and thoughts and encouraged the patients to write about them. The patients were instructed to describe the traumatic event thoroughly, including their intimate fears and thoughts concerning the traumatic experience. To increase the effect of the exposure, patients were asked to write in the first person and in present tense and to give detailed descriptions of all sensory details they experienced during the traumatic event,

including olfactory, visual, and auditory stimuli. Participants were explicitly asked not to concentrate on style, grammar, spelling, or the chronological order of their essays. The therapists checked whether patients explicitly addressed the traumatic event as described above and, if needed, supported the patient to address the avoided features more forcefully.

Second Phase: Cognitive Restructuring

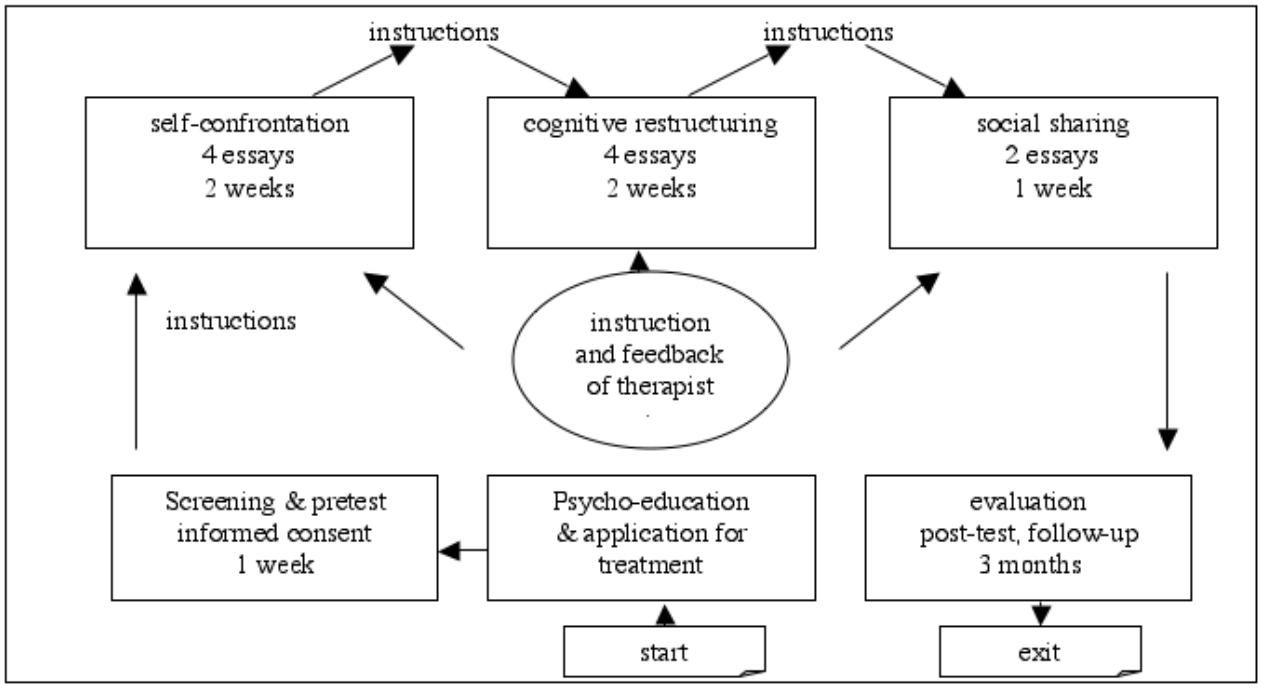
During the second phase, patients received psychoeducation on the principles of cognitive restructuring. The goal of this phase was to form a new perspective of the traumatic event and to regain a sense of control. Participants wrote a supportive letter to an imaginary friend who had been through the same experience. In this letter, the patient was instructed to reflect on the addressee's feelings of guilt and shame, challenge dysfunctional automatic thinking and behavior patterns, and correct unrealistic assumptions. Furthermore, patients were encouraged to consider potentially positive consequences of the traumatic event for that person's life and the lessons to be learned from it.

Third Phase: Social Sharing and Farewell Ritual

During the third phase, patients received psychoeducation on the positive effects of social sharing. In a final letter, they took symbolic leave of the traumatic event. Patients summarized what happened to them, reflected on the therapeutic process, and described how they were going to cope now and in the future. Patients could address the letter either to themselves, to a close friend, or another significant person involved in the traumatic event. The letter did not ultimately have to be sent.

At the beginning of each writing phase, patients proposed individual timetables as to when they planned to write. Halfway through and at the end of each treatment phase, patients received feedback and further writing instructions, which consisted of standard instructions and standard feedback tailored to patients' specific needs. Important aspects of this feedback were recognition and reinforcement of the patients' independent work, positive feedback, motivation and unconditional support, as well as frequent summaries and encouragement of patients to voice questions and doubts. An overview of the Interapy procedure is given in Figure 1.

Figure 1. Overview of the Interapy procedure



The Therapeutic Alliance

The WAI [20] is a transtheoretical measure of the working alliance that was designed to apply to diverse therapeutic orientations and modalities and is one of the questionnaires most frequently used to measure the working alliance [28]. In this study, the WAI-S [29] was used. Busseri and Tyler [28] have shown that the two versions correlate highly in terms of their psychometric and predictive qualities and are thus interchangeable. The WAI-S is a 12-item, self-report questionnaire consisting of three subscales designed to assess three primary components of the working alliance: (1) how closely client and therapist agree on and are mutually engaged in the goals of treatment (goal subscale reliability coefficient in this study:α = .79), (2) how closely client and therapist agree on how to reach the treatment goals (task subscale reliability coefficient in this study:α = .70), and (3) the degree of mutual trust, acceptance, and confidence between client and therapist (bond subscale reliability coefficient in this study:α = .75). The composite score (reliability coefficient in this study: = .83) is used as a global measurement of working alliance. Respondents were asked to rate each statement on a 7-point Likert scale ranging from 1 (never) to 7 (always).

Outcome Measurements

To assess posttraumatic stress, the revised version of the Impact of Event Scale (IES-R) [30] was used. The scale consists of 22 items constituting the subscales of (1) intrusions, (2) avoidance,

and (3) hyperarousal, the three main characteristics of psychological dysfunction after a traumatic life event. Participants were asked to indicate the frequency of each symptom over the past 7 days on a 4-point Likert scale. The presence of a PTSD diagnosis was assessed using the cut-off score proposed by Neal et al [31]. They found that an optimum cut-off score for the IES (which comprises the avoidance and intrusion subscales) of 35.0 produced the highest predictive value.

To measure depression and anxiety, the appropriate subscales of the short form of the Symptom Checklist (SCL-90), the Brief Symptom Inventory (BSI) [32], were used to measure the effects of treatment on psychological dysfunction in dimensions related to symptoms of posttraumatic stress. The two subscales consist of six items each. Each item is rated on a 5-point Likert scale (0 = not at all, 4 = extremely).

Results

Statistical analyses were only performed on the data of the 48 participants in the treatment group. Participants in this group ranged from 18 to 68 years of age, with an average age of 35 years; 92% were female; 55% had a university degree, and a further 14% had a secondary school leaving certificate. The most frequently reported traumatic event was the sudden or violent death of a close person (40%); 38% of the patients reported sexual abuse, incest, or rape; 10% were crime victims. On average, the traumatic event had occurred 3.5 years previously (range 2-696 months).

Scores on the IES-R indicated that the 48 participants suffered greatly. The mean scores on the intrusions (mean = 23.1; SD = 6.5) and avoidance (mean = 19.4; SD = 9.9) subscales were in the upper regions of the norm table for Dutch PTSD patients [33].

Dropouts

Those who terminated the treatment early (dropout 17%, N = 8) were compared with those who completed the program in terms of demographic variables. Chi-square analyses failed to reveal any significant differences between dropouts and completers in terms of gender, education level, or marital status, and a *t* test showed no significant differences in terms of age or years since the trauma. We also used *t* tests to assess differences between dropouts and completers in pretreatment psychopathology: no significant differences were found for depression (BSI) (t_{46} = .78, P = .44), anxiety (BSI) (t_{46} = .84, P = .41), posttraumatic symptoms (IES-R) (t_{46} = −1,077, P = .29), or WAI-S (t_{41} = −.639, P = .53). Note, however, that WAI-S scores were only available for three dropouts. The other dropouts terminated therapy before the WAI-S was administered.

Patients' Pretherapy Status and Ratings of Treatment Relationship

Focusing on the 40 patients who completed the course of therapy, zero-order Pearson correlations were used to assess the relations between variables (Table 1). Bivariate analysis of relationships between pretreatment psychopathology and working alliance scores revealed no significant pattern of relationships but showed a tendency of an inverse relationship.

Table 1. Means, standard deviations, and correlations of patients' scores on the WAI-S (at 4th session‡) and initial symptoms (at 1st session†) (N = 40)

	Mean	SD	Goal ‡	Task ‡	Bond ‡	Composite ‡
WAI-S ‡						
Goal	5.8	.77	1			
Task	5.7	.80	.90	1		
Bond	6.2	.75	.31	.15	1	
Composite	5.8	.62	.90	.83	.64	1
IES-R †						
Intrusions	24.4	6.2	.12	.04	−.11	.04
Avoidance	18.9	10.2	−.12	.11	−.35*	−.19
Hyperarousal	21.6	6.7	.09	.02	−.20	−.08
BSI †						
Anxiety	9.5	3.2	−.16	−.11	−.34*	−.26
Depression	10.4	4.0	.09	.01	−.13	−.04

*P < .05

†at 1st session

‡at 4th session

Association of the Working Alliance With Therapy Outcome

As indices of client outcome, the residual gain scores on each subscale of the self-report measures (BSI, IES-R) were calculated. Each participant's residual gain score at each posttreatment assessment point was the deviation of the posttreatment score on that measure from the pretreatment assessment. Residual gain scores were reversed as appropriate so that higher scores indicate greater improvement (ie, greater reduction in psychopathology). These residual gain scores across the patients were correlated with their scores on the WAI-S. Table 2 shows partial correlations between the patients' scores on the WAI-S (subscales and composite) and their posttreatment scores on target variables (BSI, IES-R) after partialing out initial symptom levels.

Table 2. Means, standard deviations, and correlations of the WAI-S patient and therapist ratings and residual gain (N = 40)

			WAI-S Patient Ratings				WAI-S Therapist Ratings			
	Mean	SD	Goal	Task	Bond	Composite	Goal	Task	Bond	Composite
IES-R residual gain										
Intrusions	13.0	9.4	.15	.17	.01	.16	.08	.09	.05	.08
Avoidance	11.8	10.6	.22	.22	−.12	.13	.19	.25	−.22	.08
Hyperarousal	13.0	9.0	.09	.09	.13	.15	.03	−.08	−.05	−.04
BSI residual gain										
Anxiety	4.9	4.2	.27	.24	.19	.33*	.30	.27	.09	.25
Depression	6.0	4.4	.27	.29	−.03	.21	.17	.21	.13	.20

*$P < .05$

Positive correlations were found between the patients' ratings of the working alliance and therapy outcome. However, with the exception of the relation between the WAI-S composite score and anxiety ($r = .33$, N = 40, $P = .04$), these correlations did not reach statistical significance. For the most part, positive correlations were also found between the therapists' ratings of the working alliance and the outcome, although these did not reach statistical significance either. Multiple regression analyses were used to further explore possible mediator or suppressor effects of the working alliance on outcome variables (residual change in IES-R composite score and residual change in BSI anxiety and BSI depression). Results revealed that the working alliance, as rated by patients, did not exert a significant direct influence on posttraumatic symptoms (adjusted $R^2 = -.026$, $F_{1,38} = .007$, $P = .93$); depression (adjusted $R^2 = -.026$, $F_{1,38} = .005$, $P = .94$); or anxiety (adjusted $R^2 = -.017$, $F_{1,38} = .358$, $P = .55$).

Discussion

The aim of this study was to investigate the quality and the possible influence of an Internet-based therapeutic relationship on treatment outcome. To our knowledge, this was the first study in which the effects of the working alliance have been systematically evaluated in an Internet-based therapy approach. Bearing in mind that the generalizability of our findings is limited by the small sample size and the narrow diagnostic range of clients, we now turn to the research questions raised above.

Does Baseline Psychopathology Predict the Quality of the Working Alliance?

No significant relationship was detected between the severity of pretreatment psychopathology and the working alliance rating. However, a tendency of an overall inverse relationship was observed, indicating that patients who experienced more severe symptoms at the beginning of the treatment tended to have a less positive relationship with their therapist. This would be in line with previous research by Taft et al [34], who found a significant inverse correlation ($r = -.31$) between psychopathology and early working alliance ratings in face-to-face therapy.

Is the Quality of the Therapeutic Alliance Linked to Treatment Outcome?

The results failed to confirm the hypothesis that a strong working alliance early in treatment would predict positive psychological changes later in treatment. However, almost all of the correlations were positive, indicating that residual gains on outcome measures were associated with higher rather than lower mean WAI-S scores, except in the relation between working alliance and anxiety. The finding that the WAI-S failed to predict therapy outcome in our sample stands in marked contrast to the findings for most face-to-face studies. This discrepancy may be attributable to a number of factors. One explanation for the lack of effect may be the almost uniformly high levels of alliance ratings (ie, restricted range) obtained in this study, perhaps due to the self-selected sample. Most of the patients were recruited through the Internet, which suggests that they were already comfortable with this medium. Research has shown that computer experience influences the way people judge Internet-based contact. In their study, Mallen et al [1] showed that the more familiar participants were with Internet-based contact, the more positively they judged that contact to be.

Another possible reason for the failure to find more substantial relationships between the quality of the working alliance and treatment outcome has been proposed by Stiles et al [11], who found great variability in the correlation with outcome measures

taken at different stages in the therapy. They suggest that this might explain why various studies in which the working alliance was only measured on a single occasion produced inconsistent alliance/outcome correlations. This line of reasoning suggests that the question is not whether the working alliance is more important in a particular type of therapy, but rather whether the alliance is being measured in a way that is appropriate to that particular therapy. The time of administration of the WAI (in terms of the number of sessions) has been found to influence the rating of the working alliance [5,29]. It has also been suggested that treatment outcome may be particularly well predicted by the quality of the working alliance as measured in early sessions [5,8,11]. As is standard practice in face-to-face studies, the therapeutic alliance was assessed early in the therapeutic process in the present study, after the fourth writing session [28]. At that point, however, there had been only three therapist/client contacts, which may not in fact have been sufficient to evaluate the therapeutic alliance in online therapy. It could be that, given the different conditions under which the working alliance develops in Internet-based treatment approaches, administering the WAI later on in the therapy might yield more accurate measurements.

Alternatively, although the alliance has been shown to predict the outcome of other modes of delivery, it may not be a crucial factor in facilitating positive psychological change in Internet-based manualized therapies. The treatment applied in this study incorporates principles derived from CBT, with standardized instructions and a fixed treatment manual, and focuses on client empowerment and self-efficacy. It may be that the nonspecific factor of the therapeutic relationship played a less important role than it does in less structured face-to-face therapy.

Can a Positive and Stable Relationship Be Established Through the Internet?

Patients reported high levels of therapeutic alliance early in treatment. The patients' ratings of the therapeutic relationship in our study were even higher than in face-to-face studies. Hersoug et al [35] administered the WAI to 270 patients with multiple clinical disorders in the third session of a conventional face-to-face therapy approach. Compared to the mean composite score in their study (mean = 4.94; SD = 1.08), our patients' ratings of the Internet-based relationship were more than one standard deviation higher (mean = 5.8; SD = 6.2), as shown in Table 1. The patients' positive evaluation of the therapeutic relationship indicates that a therapeutic alliance can be established through the Internet. Furthermore, a strong working alliance can be expected to promote treatment adherence as assessed by factors such as dropout rates. Given that trauma victims have been shown to have compliance problems [36,37] and high dropout rates (up to 28%) [38], the high WAI-S ratings and the relatively low dropout rate (17%) in this study give reason to conclude that it was indeed possible to develop a positive and stable therapeutic relationship through the Internet. It must be noted, however, that the therapeutic contact as performed during the Interapy treatment is rather exceptional. The intensity of individualized support and regular personal interaction differs markedly from online approaches where online personalized communication is rather uncommon.

Limitations

The following limitations necessitate caution in the interpretation of our results. First, the modest sample size may have provided insufficient power to uncover the complex interplay of the online working alliance and psychopathology measures.

Second, only 17% of applicants could be included in the study, which might limit the external validity of the present findings. The same applies to the specific sample of trauma victims. Trauma survivors have been noted by many clinicians to have difficulty in tolerating the interpersonal nature of therapy, particularly "the [need] to trust another person with his or her pain" [39] (p. 538). Given that trauma victims are especially prone to feelings of guilt and shame, they might be especially drawn to the medium of the Internet, where visual anonymity enables them to disclose painful and shameful details more easily than in face-to-face settings. Extending this research paradigm to clinical samples other than trauma victims could help to clarify this relationship.

Third, further research efforts should be initiated to address the possibility that the results are only valid for users who are already comfortable with the Internet due to self-selection. A direct comparison of an online intervention and a face-to-face intervention as a randomized controlled trial would be indicated to investigate how the text-based bond formed in online therapy compares and contrasts with the in-person therapeutic alliance.

Limitations notwithstanding, the findings presented here are of interest because they indicate that a stable and positive relationship can be established online, although the quality of the relationship does not predict treatment outcome. The rapid growth of Internet-based treatment approaches makes it likely that online therapies will become an enduring component of the psychotherapeutic landscape. One line of future research will be to identify predictors of a positive therapeutic relationship. A major challenge when building online relationships is to become aware of the nuances in the written language used in this context [12], which has accents, ambiguities, and individual styles, as well as the use of emoticons (emotion + icon; eg, a happy face ☺). Clinicians who work online should be given clinical training focusing on features of written communication. In addition, further work is needed to determine whether the role of the working alliance differs as a function of the mode of delivery, and to disentangle the relationships between the therapeutic alliance, specific cognitive-behavioral techniques, and treatment outcome. It also remains to be seen whether working alliance scores will predict long-term reductions in psychopathology rather than focusing on short-term changes in psychological functioning, as was the case in this study.

Acknowledgments

The authors would like to thank Alfred Lange, Pieter van Hoogstraten, and other team members of Interapy for their intellectual and logistic support and for helpful suggestions at various stages of our treatment study. We would also like to thank our colleague Birgit Wagner for serving as a therapist and also providing essential suggestions at various stages of our treatment study. This study was co-funded by the German organization for crime victims, Weisser Ring.

Conflicts of Interest

None declared.

References

1. Mallen MJ, Day SX, Green MA. Online versus face-to-face conversations: an examination of relational and discourse variables. Psychother-Theor Res 2003;40:155-163.
2. Norcross JC, Hedges M, Prochaska JO. The face of 2010: a Delphi poll on the future of psychotherapy. Prof Psychol-Res Pr 2002;33(3):316-322. [doi: 10.1037/0735-7028.33.3.316]
3. Knaevelsrud C, Jager J, Maercker A. Internet-Psychotherapie: Wirksamkeit und Besonderheiten der therapeutischen Beziehung [Internet psychotherapy: efficacy and particularities of the therapeutic relationship]. Verhaltenstherapie 2004;14(3):174-183. [doi: 10.1159/000080913]
4. Rochlen AB, Zack JS, Speyer C. Online therapy: review of relevant definitions, debates, and current empirical support. J Clin Psychol 2004 Mar;60(3):269-283. [Medline: 14981791] [doi: 10.1002/jclp.10263]
5. Horvath AO, Luborsky L. The role of the therapeutic alliance in psychotherapy. J Consult Clin Psychol 1993 Aug;61(4):561-573. [Medline: 8370852] [doi: 10.1037/0022-006X.61.4.561]
6. Gaston L. The concept of the alliance and its role in psychotherapy: theoretical and empirical considerations. Psychother 2004;27:143-153.
7. Horvath AO, Greenberg LS. The working alliance: theory, research, practice. New York: Wiley; 1994.
8. Horvath AO, Symonds BD. Relation between working alliance and outcome in psychotherapy: a meta-analysis. J Couns Psychol 1991;38(2):139-149. [doi: 10.1037/0022-0167.38.2.139]
9. Krupnick JL, Sotsky SM, Simmens S, Moyer J, Elkin I, Watkins J, et al. The role of the therapeutic alliance in psychotherapy and pharmacotherapy outcome: findings in the National Institute of Mental Health Treatment of Depression Collaborative Research Program. J Consult Clin Psychol 1996 Jun;64(3):532-539. [Medline: 8698947] [doi: 10.1037/0022-006X.64.3.532]
10. Martin DJ, Garske JP, Davis MK. Relation of the therapeutic alliance with outcome and other variables: a meta-analytic review. J Consult Clin Psychol 2000 Jun;68(3):438-450. [Medline: 10883561] [doi: 10.1037/0022-006X.68.3.438]
11. Stiles WB, Agnew-Davies R, Hardy GE, Barkham M, Shapiro DA. Relations of the alliance with psychotherapy outcome: findings in the Second Sheffield Psychotherapy Project. J Consult Clin Psychol 1998 Oct;66(5):791-802. [Medline: 9803698] [doi: 10.1037/0022-006X.66.5.791]
12. Skarderud F. Sh@me in cyberspace: relationships without races: the emedia and eating disorders. Eur Eat Disord Rev 2003;11(3):155-169. [doi: 10.1002/erv.523]
13. Slouka M. War of the worlds: cyberspace and the high-tech assault on reality. New York: Basic Books; 1995.
14. Kraut R, Patterson M, Lundmark V, Kiesler S, Mukopadhyay T, Scherlis W. Internet paradox. A social technology that reduces social involvement and psychological well-being? Am Psychol 1998 Sep;53(9):1017-1031. [Medline: 9841579] [doi: 10.1037/0003-066X.53.9.1017]
15. Walter JB. Computer-mediated communication: impersonal, interpersonal, and hyperpersonal interaction. Commun Res 1996;23(1):3-43. [doi: 10.1111/j.1468-2958.1996.tb00384.x]
16. Whitty M, Gavin J. Age/sex/location: uncovering the social cues in the development of online relationships. Cyberpsychol Behav 2001 Oct;4(5):623-630. [Medline: 11725656] [doi: 10.1089/109493101753235223]
17. Suler J. The online disinhibition effect. Cyberpsychol Behav 2004 Jun;7(3):321-326. [Medline: 15257832] [doi: 10.1089/1094931041291295]
18. Utz S. Social information processing in MUDs: the development of friendships in virtual worlds. Journal of Online Behavior. 2000. (1) URL: http://www.behavior.net/JOB/v1n1/utz.html [accessed 2006 Dec 2] [WebCite Cache ID 5FPX1FKsU]
19. Cook JE, Doyle C. Working alliance in online therapy as compared to face-to-face therapy: preliminary results. Cyberpsychol Behav 2002 Apr;5(2):95-105. [Medline: 12025884] [doi: 10.1089/109493102753770480]
20. Horvath AO, Greenberg LS. Development and validation of the Working Alliance Inventory. J Couns Psychol 1989;36(2):223-233. [doi: 10.1037/0022-0167.36.2.223]
21. Lange A, van de Ven JP, Schrieken B, Smit M. Interapy burn-out; preventie en behandeling van burn-out via Internet [Interapy burnout; prevention and treatment of burnout through the Internet]. Directieve Therapie 2003;23:121-145.
22. Cohen E, Kerr BA. Computer-mediated counseling: an empirical study on a new mental health treatment. Comput Hum Serv 1998;15:77-81.

23. Lange A, Schoutrop M, Schrieken B, van de Ven JP. Interapy: a model for therapeutic writing through the Internet. In: Lepore SJ, Smyth JM, editors. The writing cure: how expressive writing promotes health and emotional well-being. Washington, DC: APA Books; 2002:215-238.
24. ; American Psychiatric Association. Diagnostic and Statistical Manual of Mental Disorder DSM IV. Washington, DC: American Psychiatric Association; 1994.
25. Nijenhuis ER, Spinhoven P, van Dyck R, van der Hart O, Vanderlinden J. The development of the somatoform dissociation questionnaire (SDQ-5) as a screening instrument for dissociative disorders. Acta Psychiatr Scand 1997 Nov;96(5):311-318. [Medline: 9395146]
26. Lange A, Schrieken B, Blankers M, van de Ven JP, Slot M. Constructie en validatie van de Gewaarwordingenlijst: een hulpmiddelbij het signaleren van een verhoogde kans op psychosen [Construction and validation of the Screening Device for Psychotic Disorder]. Directieve Therapie 2000;20:162-173.
27. Arnoldi J, van de Ven JP, Schrieken B, Lange A. De Risico Taxatielijst- eigenschappen van een kort screeningsintrument [Suicide Risk Assessment - a short screening tool]. Unpublished manuscript University of Amsterdam; 2000.
28. Busseri MA, Tyler JD. Interchangeability of the Working Alliance Inventory and Working Alliance Inventory, Short Form. Psychol Assess 2003 Jun;15(2):193-197. [Medline: 12847779] [doi: 10.1037/1040-3590.15.2.193]
29. Tracey TJ, Kokotovic AM. Factor structure of the Working Alliance Inventory. Psychol Assess 1989;1(3):207-210. [doi: 10.1037/1040-3590.1.3.207]
30. Weiss DS, Marmar CR. The impact of event scale-revised. In: Wilson JP, Keane TM, editors. Assessing psychological trauma and PTSD. New York: Guilford Press; 1997:399-411.
31. Neal LA, Busuttil W, Rollins J, Herepath R, Strike P, Turnbull G. Convergent validity of measures of post-traumatic stress disorder in a mixed military and civilian population. J Trauma Stress 1994 Jul;7(3):447-455. [Medline: 8087405] [doi: 10.1007/BF02102789]
32. Derogatis LR. BSI: administration, scoring, and procedures manual-II. 2nd edition. Baltimore, MD: Clinical Psychometric Research, Inc; 1992.
33. Kleber RJ, Brom D. Traumatische ervaringen, gevolgen enverwerking [Traumatic events, consequences and processing]. Lisse, The Netherlands: Swets & Zeitlinger; 1986.
34. Taft CT, Murphy CM, Musser PH, Remington NA. Personality, interpersonal, and motivational predictors of the working alliance in group cognitive-behavioral therapy for partner violent men. J Consult Clin Psychol 2004 Apr;72(2):349-354. [Medline: 15065967] [doi: 10.1037/0022-006X.72.2.349]
35. Hersoug AG, Høglend P, Monsen JT, Havik OE. Quality of working alliance in psychotherapy: therapist variables and patient/therapist similarity as predictors. J Psychother Pract Res 2001 Dec;10(4):205-216 [FREE Full text] [Medline: 11696646]
36. Maercker A. Therapie der posttraumatischen Belastungsstörung [Therapy of posttraumatic stress disorder].9th edition. Berlin, Germany: Springer; 2003.
37. Scott MJ, Stradling SG. Client compliance with exposure treatments for posttraumatic stress disorder. J Trauma Stress 1997 Jul;10(3):523-526. [Medline: 9246657] [doi: 10.1023/A:1024805807952]
38. Foa EB, Rothbaum BO, Riggs DS, Murdock TB. Treatment of posttraumatic stress disorder in rape victims: a comparison between cognitive-behavioral procedures and counseling. J Consult Clin Psychol 1991 Oct;59(5):715-723. [Medline: 1955605] [doi: 10.1037/0022-006X.59.5.715]
39. Turner SW, McFarlane AC, van der Kolk BA. The therapeutic environment and new explorations in the treatment of posttraumatic stress disorder. In: van der Kolk BA, McFarlane AC, Weisaeth L, editors.Traumatic stress: the effects of overwhelming experience on mind, body, and society. New York: Guilford Press; 1996:538.

Abbreviations

BSI: Brief Symptom Inventory
CBT: cognitive-behavioral therapy
DSM-IV: Diagnostic and Statistical Manual of Mental Disorders, Fourth Edition
IES-R: revised version of the Impact of Event Scale
PTSD: posttraumatic stress disorder
SCL-90: Symptom Checklist
SDQ-5: Somatoform Dissociation Questionnaire
SRA: Suicide Risk Assessment
WAI: Working Alliance Inventory
WAI-S: short version of the WAI

submitted 27.04.06; peer-reviewed by G Andersson; comments to author 08.08.06; revised version received 04.09.06; accepted 03.11.06; published 19.12.06.

Please cite as:
Knaevelsrud C, Maercker A
Does the Quality of the Working Alliance Predict Treatment Outcome in Online Psychotherapy for Traumatized Patients?
J Med Internet Res 2006;8(4):e31
URL: http://www.jmir.org/2006/4/e31/
doi:10.2196/jmir.8.4.e31
PMID:17213049

Publisher:
JMIR Publications
130 Queens Quay East.
Toronto, ON, M5A 3Y5
Phone: (+1) 416-583-2040
Email: support@jmir.org