Health Informatics Journal

Volume 24 Number 1 March 2018

ISSN 1460-4582
journals.sagepub.com/home/jhi

Health Informatics Journal

Article

Health Informatics Journal
2018, Vol. 24(1) 3–13
© The Author(s) 2016
Reprints and permissions:
sagepub.co.uk/journalsPermissions.nav
DOI: 10.1177/1460458216656472
journals.sagepub.com/home/jhi
SAGE

A randomized controlled pilot study of CBT-I Coach: Feasibility, acceptability, and potential impact of a mobile phone application for patients in cognitive behavioral therapy for insomnia

Erin Koffel
Minneapolis Veteran Affairs Health Care System, USA; University of Minnesota Medical School, USA

Eric Kuhn
National Center for PTSD (NCPTSD), Dissemination and Training (D&T) Division, USA; Department of Veterans Affairs Palo Alto Health Care System (VAPAHCS), USA

Napoleon Petsoulis
Widener University, USA

Christopher R Erbes
Minneapolis Veteran Affairs Health Care System, USA; University of Minnesota Medical School, USA

Samantha Anders
Hennepin County Medical Center, USA

Julia E Hoffman and Josef I Ruzek
National Center for PTSD (NCPTSD), Dissemination and Training (D&T) Division, USA; Department of Veterans Affairs Palo Alto Health Care System (VAPAHCS), USA

Melissa A Polusny
Minneapolis Veteran Affairs Health Care System, USA; University of Minnesota Medical School, USA

Corresponding author:
Erin Koffel, Center for Chronic Disease Outcomes Research, Minneapolis Veteran Affairs Health Care System, One Veterans Drive, Minneapolis, MN 55417, USA.
Email: Erin.Koffel@va.gov

Abstract

There has been growing interest in utilizing mobile phone applications (apps) to enhance traditional psychotherapy. Previous research has suggested that apps may facilitate patients' completion of cognitive behavioral therapy for insomnia (CBT-I) tasks and potentially increase adherence. This randomized clinical trial pilot study ($n=18$) sought to examine the feasibility, acceptability, and potential impact on adherence and sleep outcomes related to CBT-I Coach use. All participants were engaged in CBT-I, with one group receiving the app as a supplement and one non-app group. We found that patients consistently used the app as intended, particularly the sleep diary and reminder functions. They reported that it was highly acceptable to use. Importantly, the app did not compromise or undermine benefits of cognitive behavioral therapy for insomnia and patients in both groups had significantly improved sleep outcomes following treatment.

Keywords

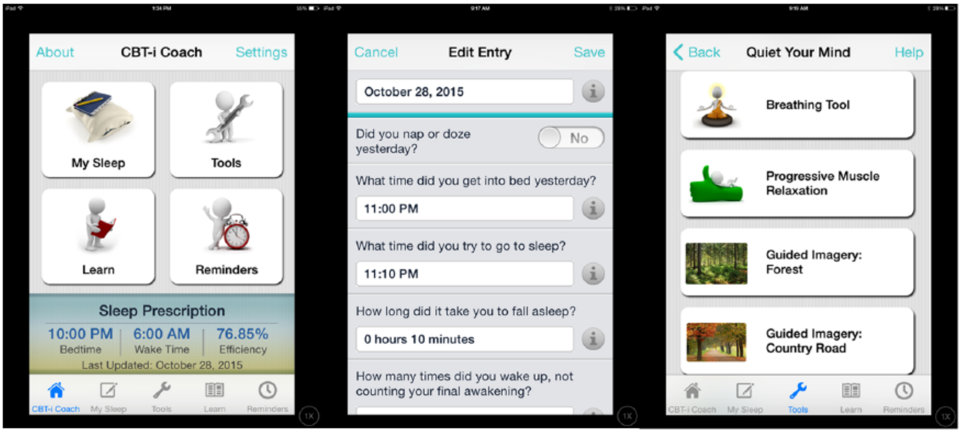
cognitive behavior therapy, computer applications, insomnia, mobile devices, psychotherapy

Chronic insomnia affects more than 1 in 10 people.[1] Rates are even higher among veterans using Veterans Health Administration (VHA) services.[2] Cognitive behavioral therapy for insomnia (CBT-I) is an effective treatment for insomnia[3–7] with superior long-term efficacy[8] and fewer risks compared to hypnotic medications.[9] As such it is widely recommended as a first-line treatment for insomnia.[1]

The VHA has an ongoing national dissemination initiative to train its licensed mental health care providers (i.e. non-sleep specialists) to deliver CBT-I.[10] Patient outcomes from this training program are comparable to those of published research trials.[11,12] However, patient non-adherence to treatment recommendations (e.g. going to bed only when sleepy and limiting naps) may be undermining the benefit for some patients.

Mobile applications (apps) utilized with smartphones have the potential to improve traditional psychotherapy by enhancing access to psychoeducation and psychotherapy skills (e.g. relaxation techniques), facilitating monitoring of symptoms and outcomes, and assisting with relapse prevention.[13,14] Emerging literature is beginning to accumulate showing the promise of apps for depression, stress, psychosis, eating disorders, and substance use[15–19] and integrating mental health apps into treatment within a VA setting shows initial promise.[20] Although there is increasing availability of consumer sleep technologies for sleep improvement, many of the existing apps focus on monitoring sleep and providing sleep education, rather than assisting with active therapy.[21–23]

The VHA National Center for Post-Traumatic Stress Disorder (PTSD), in partnership with Stanford University School of Medicine and the Department of Defense's National Center for Telehealth and Technology, built CBT-I Coach, a patient app designed to be used as an adjunct to face-to-face CBT-I. It was released for both iOS[24] and Android[25] mobile devices in 2013. It was specifically designed to help facilitate patients' completion of CBT-I tasks, potentially increasing patient adherence to the protocol.[26] The design and content of the app has been described in detail in previous publications.[26,27] Briefly, the app provides education about sleep processes, developing positive sleep routines, and improving sleep environments. Key features of the app include a sleep diary to record daily sleep variables (see Figure 1), ability to update a sleep prescription (recommended bedtime and wake time) in consultation with CBT-I providers, tools and guided exercises for quieting the mind (see Figure 1), education about sleep and sleep-health behaviors, reminder functions and alarms to help change sleep habits (e.g. reminders for when to stop caffeine intake, start wind-down time, and alarms for prescribed bedtime and wake time). See Figure 1 for a screenshot of the home page of the app.



Sleep Prescription
10:00 PM Bedtime | 6:00 AM Wake Time | 76.85% Efficiency
Last Updated: October 28, 2015

October 28, 2015

Did you nap or doze yesterday? No

What time did you get into bed yesterday? 11:00 PM

What time did you try to go to sleep? 11:10 PM

How long did it take you to fall asleep? 0 hours 10 minutes

How many times did you wake up, not counting your final awakening?

Figure 1. Example screenshots from CBT-I Coach.

In a recent survey, VA CBT-I clinicians indicated that they believe this app may improve care and increase adherence,[26] and an initial feasibility study investigating the use of the app among patients engaged in a cannabis cessation attempt indicated daily engagement with the app.[27] However, it is uncertain if integrating CBT-I Coach into the existing VA CBT-I protocol is feasible in terms of whether patients will consistently use it as intended and if they will find it acceptable to use while engaged in this therapy. It is also unknown if CBT-I Coach will actually improve patient adherence to the therapy and, importantly, whether using the app will compromise or undermine benefits of CBT-I, a well-established evidence-based treatment, on sleep-related outcomes. Therefore, the current study sought to examine the feasibility and acceptability of CBT-I Coach, as well as explore the potential impact of this device on adherence and sleep outcomes. The first-author, in conjunction with the developers of the app (second, sixth, and seventh authors) utilized the existing app in a sample of patients receiving CBT-I in a clinical setting. We hypothesized that participants randomized to CBT-I with the app and without the app would report significant improvements in sleep, but that the app group would have higher adherence than the non-app group.

Methods

Participants

This study was approved by the Veterans Affairs Institutional Review Board. Forty-one consecutive referrals for CBT-I at a Midwestern VA Medical Center were pre-screened following CBT-I intake assessments by project clinicians, with further screening via phone by the project coordinator. Inclusion criteria included commitment to begin CBT-I, ownership of a smartphone, and willingness to use CBT-I Coach. Exclusion criteria included moderate or greater suicidal or homicidal ideation, significant alcohol or drug use, and active psychotic symptoms. Patients not meeting these criteria where referred elsewhere or underwent CBT-I outside of the study.

Figure 2 presents the flow of participants through the study from March 2014 to November 2014. Twenty-three potential participants were excluded, most commonly for not owning a smartphone ($n = 9$) and not imminently planning to begin CBT-I ($n = 8$). Two participants declined to participate due to lack of interest. After obtaining informed consent, 18 participants were randomized equally to either CBT-I plus CBT-I Coach (app group) or CBT-I without the app (non-app

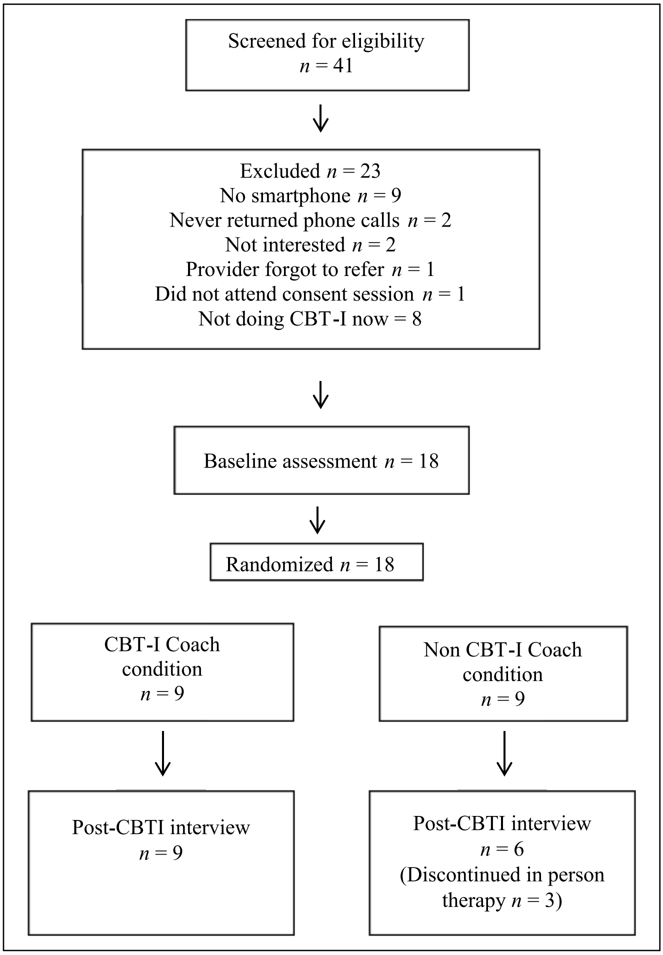


Figure 2. Flow of participants through the study.

group) by the project coordinator, utilizing a computerized random number generator. Table 1 presents baseline characteristics. The average age was 48.50 years (standard deviation (SD) = 14.93), and participants were primarily White males with Android smartphones. There were no significant differences between the two conditions on baseline insomnia symptom levels or demographics, with the exception of marital status ($p < .05$), where those in the app group were more likely to be married compared to the non-app group (55.56% vs 11.11%). On average, participants reported clinical levels of insomnia. Three participants dropped out of the non-app group (two lost interest in treatment and dropped out after three sessions, one moved and dropped out after two sessions); none dropped out of the app group. There were no significant differences between retained participants and those who dropped out on baseline insomnia symptom levels or demographics.

Procedure

Participants completed semi-structured interviews at baseline and following treatment, as well as self-report measures prior to each CBT-I treatment session. Participants received a US$25 gift card

Table 1. Baseline characteristics.

	Coach	Non-coach
	(N = 9)	(N = 9)
Demographics		
Age, mean (SD)	50.11 (15.74)	46.89 (14.85)
Men (%)	6 (66.67)	5 (55.56)
White race (%)	8 (88.89)	9 (100.00)
Married (%)	5 (55.56)	1 (11.11)*
Education > High School (%)	8 (88.89)	7 (77.78)
Sleep		
Insomnia Severity Index (SD)[a]	19.22 (3.70)	20.38 (5.55)
Phone		
Phone type (%)		
iPhone	2 (22.22)	3 (33.33)
Android	7 (77.78)	6 (66.67)

SD: standard deviation.

[a] Based on n = 8 in non-coach due to missing data.

*Significant χ^2 difference at $p \leq .05$. Insomnia Severity Index ≥ 15 indicates clinical levels of insomnia.

for completing the baseline assessment and a US$50 gift card for the final assessment. Pre-treatment interviews were identical for both app and non-app groups and were used to obtain demographic information and baseline experience with apps, including questions about phone usage and comfort, and app usage. Following the initial interview, participants assigned to the app group were shown how to download and use the app. The app participants used the app in conjunction with the standard CBT-I procedures, including using it to complete sleep diaries. Non-app group participants completed CBT-I according to standard procedures without CBT-I Coach.

In order to explore the impact of the app on adherence to treatment recommendations and sleep outcomes, participants in both groups reported the number of days they completed homework, amount of time spent on homework each week, and completed an insomnia questionnaire before each therapy session. Therapists completed a measure of patient adherence to treatment recommendations at the end of each session for all participants. At post-treatment, app group participants completed semi-structured interviews focused on their use and engagement with the app, including their impressions of CBT-I Coach, which elements they used, barriers to app usage, perceived value of the app, and potential enhancements of the app. The non-app group participated in a semi-structured interview during which they were shown the app and asked for their thoughts on integrating the app into CBT-I and suggestions for enhancements.

Treatment and therapists

Participants in both conditions completed CBT-I with one of two VA clinical psychologists who completed the VHA CBT-I training initiative.[10,11] Treatment consisted of weekly 1-h individual therapy sessions based on the CBT-I manual developed by VA.[10] The protocol consists of five treatment sessions, with patients attending fewer or more sessions if clinically indicated. The average number of treatment sessions completed in this study were four (66.67% of patients), with almost a third of patients completing five sessions (27.78% of patients). The basic components of this protocol include: (1) sleep restriction, which involves limiting time in bed to consolidate sleep; (2)

stimulus control, which involves restricting the bed/bedroom to sleep; and (3) cognitive restructuring, which addresses maladaptive thoughts and beliefs about sleep.

Measures

Insomnia Severity Index. The Insomnia Severity Index (ISI)[28] is a 7-item self-report measure designed to provide a global measure of difficulties sleeping at night and daytime impairment and was completed by participants during each session. Respondents rate statements using 4-point scales from 0 to 4 (response options differ by item) and items are summed to provide a total score. Scores of 0–7 indicate no clinical insomnia, 8–14 indicate subthreshold insomnia, 15–21 indicate moderate insomnia, and 22–28 indicate severe insomnia. This instrument has adequate psychometric properties, including internal reliability (coefficient alphas ranging from .76 to .78) and concurrent validity with sleep diaries and polysomnography.[28]

Adherence scale. Adherence to CBT-I recommendations was measured using the Patient Adherence Form that was created for the VA CBT-I Training Program.[10] Starting at the end of session two, therapists rated the extent to which participants followed six specific recommendations (e.g. adhering to recommended bedtime and wake time, limiting naps, and scheduling worry time) on a scale from 1 (*no adherence*) to 6 (*complete adherence*) or not applicable (*NA*) if the recommendations had not yet been introduced during the therapy. Scores were averaged to create a total adherence score that has shown good psychometric properties.[12]

Statistical analyses

Descriptive statistics were calculated for variables related to feasibility and acceptability from the semi-structured interviews. Independent samples *t*-tests were calculated to examine variables potentially related to treatment impact, including patient adherence and homework completion. Treatment outcomes analyses were conducted using intent-to-treat for ISI scores. These analyses included all available data from 17 participants; one participant was excluded due to errors in coding outcome data. Hierarchical Linear Modeling (HLM) was conducted using SAS (PROC mixed) to determine mean values at each time point and test the effect of treatment and condition on ISI scores during five treatment sessions. Fixed effects were specified for time and condition, whereas random effects accounted for the nested nature of the data with repeated measures over time within individuals. Effect sizes were calculated using Cohen’s *d*, which represents the standardized differences between means.

Results

Feasibility and acceptability of integrating CBT-I Coach into therapy

Table 2 presents descriptive statistics of smartphone and mobile app variables for all study participants. The majority of participants indicated having their phone with them all of the time ($n=14$, 77.8%) and owning a smartphone for at least 5 years ($n=13$, 72.2%). Most participants indicated using their smartphone for calls ($n=16$, 88.9%) and apps ($n=15$, 83.33%) at least daily and most indicated being very comfortable using their smartphone generally ($n=16$, 88. 9%) and apps specifically ($n=14$, 77.8%). Very few participants had used apps for mental health or sleep ($n=2$, 11.1%) and none had used CBT-I Coach previously. Demographic variables, including age, gender, marital status, ethnicity, and education, were not significantly related to degree of comfort with apps or frequency of app use.

Table 2. Pre-treatment interview: feasibility and acceptability of mobile applications ($n = 18$).

	N (%)
How often do you have your phone with you?	
Most of the time	4 (22.22)
All of the time	14 (77.78)
Length of smartphone ownership	
≤ 4 years	5 (27.78)
5+ years	13 (72.22)
Frequency of smartphone use	
Calls	
At least daily	16 (88.89)
At least weekly	2 (11.11)
Applications	
At least daily	15 (83.33)
At least monthly	3 (16.67)
Current comfort	
Smartphone	
Very comfortable	16 (88.89)
Mostly comfortable	2 (11.11)
Applications	
Very comfortable	14 (77.78)
Mostly comfortable	1 (5.56)
Neutral	3 (16.67)
Use of smartphone applications	
Health	
Yes	9 (50.00)
No	9 (50.00)
Mental health	
Yes	2 (11.11)
No	16 (88.89)
Sleep	
Yes	2 (11.11)
No	16 (88.89)

Table 3 summarizes findings from the post-treatment semi-structured interview with the app group participants. The most commonly used element of the app was the sleep diary, followed by the educational materials, relaxation exercises, and reminders. All participants reported that the sleep diary was a helpful component (n=9, 100%; e.g. “the sleep efficiency numbers made me feel better about my sleep and decreased my anxiety about sleep. I liked being able to push enter and get the results”), followed by reminders (n=2, 22.2%; e.g. “Reminders kept me on track”). All participants indicated that they would recommend the app to family or friends. In general, feedback from participants in the app group was positive and focused on the personalized feedback provided by the app, particularly as it related to the sleep diary information (e.g. “I like the app because it gives you info on what you are doing compared to what you think you are doing”). The non-app participants also provided feedback about CBT-I Coach. One participant indicated that it would increase compliance: “I think it would be very helpful because no one loses their phone, but I lost my sleep log.” Another stated, “I think it’s a good idea, it’s a lot easier to do homework because you always have your phone with you.”

Table 3. Post-treatment interview: feasibility and acceptability of CBT-I Coach with participants in the app condition.

	$N=9$ (%)
Which elements do you use?	
Sleep diary	9 (100.0)
Learn: habits and sleep	6 (66.67)
Learn: sleep 101	5 (55.56)
Tools: quiet your mind	5 (55.56)
Reminders	5 (55.56)
I need more sleep	4 (44.44)
Tools: prevent insomnia in the future	4 (44.44)
Learn: CBT-I glossary	2 (22.22)
What part(s) of the app were most helpful?	
Sleep diary	9 (100.00)
Reminders	2 (22.22)
Relaxation	1 (11.11)
Information	1 (11.11)

CBT-I: cognitive behavioral therapy for insomnia.

Impact of CBT-I Coach on CBT-I adherence and outcomes

There were no statistically significant differences between the app and non-app group on average time spent on homework ($d=.66$ in favor of non-app group), number of days completing homework ($d=.11$ in favor of non-app group), and days completing sleep diaries ($d=.36$ in favor of app group). The effect size for the group difference on adherence scores was large ($d=.76$), favoring the app group, but it was not statistically significant ($t(16)=1.53, p=.15$).

HLM intent to treat analysis estimates for time was significant for ISI scores, $F(4, 42)=8.84$, $p<.001$, $d=1.84$ (see Figure 3). The main effect for treatment condition ($d=.21$) and the interaction of treatment time by condition ($d=.73$) were statistically non-significant. Among participants who completed treatment, four (26.7%) had no insomnia, eight (53.3%) reported subthreshold insomnia, one (6.7%) reported moderate insomnia, and two (13.3%) reported severe insomnia according to ISI scores.

Discussion

This is the first study to report the feasibility and acceptability of integrating an app with CBT-I. Previous research suggests that providers see the app as potentially improving care and increasing adherence[26] and this study suggests that integrating CBT-I Coach with individual CBT-I is highly feasible and acceptable to patients. Overall, the app was favorably received by all participants in the app group and participants in the non-app group responded favorably when they were introduced to the app at the end of therapy. The qualitative data suggest that patients were using the app as it was intended (particularly the sleep log and reminder functions) and the app improved accessibility to therapy materials.

This study is also the first to report on the impact of CBT-I Coach on process variables, including homework completion, adherence, and accessibility to therapy components. As hypothesized, CBT-I remains an effective treatment after integrating the app. Use of the app did not appear to erode or dilute the basic elements of the therapy. There was also some indication that app use

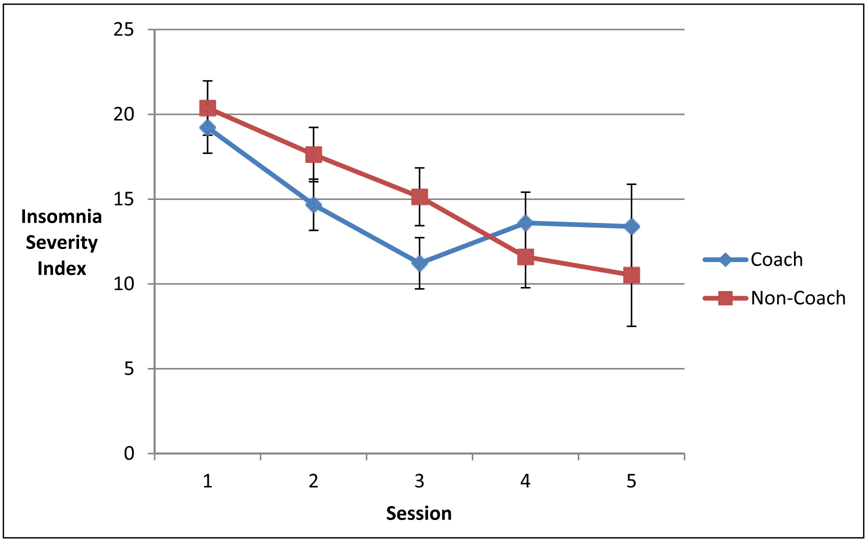


Figure 3. Hierarchical linear model predicted Insomnia Severity Index (ISI) scores at each therapy session (*n* = 17). Error bars represent standard error above and below predicted values.

related to better adherence to therapeutic recommendations. Although this finding was non-significant, the large effect size produced suggests that members of the app group are perceived as being more adherent by therapists. The study was clearly underpowered, but this potential effect represents a promising area for future research.

Although this study represents an important first step in investigating the role of mobile interventions in the treatment of insomnia, there are several limitations that should be noted. The trial was non-blinded, which may have biased adherence ratings by therapists (e.g. higher adherence scores for patients in app group). The research context may have increased homework compliance and sleep diary completion by patients in both groups (e.g. higher homework completion across all patients due to weekly monitoring, resulting in non-significant group differences). In addition, the app did not provide time-date stamped data, and so we could not confirm if participants completed the sleep diary within an hour or two of awakening as is recommended. Moreover, we were not able to collect objective information on the amount of time spent using the app, as this capacity is not built into the app.

Integrating mobile health (mHealth) into behavioral sleep treatments represents a promising area for future research, particularly within the context of stepped care models. Demand for sleep treatment often exceeds the availability of trained providers, which validates and necessitates the use of stepped care models. These models are often conceptualized in the shape of a pyramid, with the base representing low intensity, low resource treatment modalities (e.g. self-help) as an entry point, with treatments growing progressively more resource heavy and intensive as one moves to higher levels.[29,30] Apps based on CBT-I principles may represent an entry level step for some patients and pilot trials are needed with stand-alone CBT-I apps to determine treatment efficacy. In addition to potentially serving as a stand-alone treatment method, behavioral sleep treatment apps provide an efficient way to collect information (e.g. baseline sleep diary data) and deliver basic sleep education to patients while they are waiting for care. Finally, the app could potentially be utilized as a screening tool to help providers determine what level of care is necessary for an individual patient, in accordance with baseline symptom levels and initial improvement in sleep following independent app utilization; additional studies with apps are needed to determine how they can contribute to personalized treatment by matching patients with appropriate levels of care.

Acknowledgements

The views expressed in this article are those of the authors and do not reflect the official policy or position of the US Department of Veterans Affairs.

Funding

This material is the result of work supported with resources and the use of facilities at the Minneapolis VA Health Care System and the VA Palo Alto Health Care System.

References

1. National Institutes of Health. Manifestations and management of chronic insomnia in adults. *Sleep* 2005; 28: 1049–1057.
2. Mustafa M, Erokwu N, Ebose I, et al. Sleep problems and the risk for sleep disorders in an outpatient veteran population. *Sleep Breath* 2005; 9: 57–63.
3. Morin CM, Culbert JP and Schwartz SM. Nonpharmacological interventions for insomnia: a meta-analysis of treatment efficacy. *Am J Psychiatry* 1994; 151: 1172–1180.
4. Murtagh DR and Greenwood KM. Identifying effective psychological treatments for insomnia: a meta-analysis. *J Consult Clin Psychol* 1995; 63: 79–89.
5. Smith MT, Perlis ML, Park A, et al. Comparative meta-analysis of pharmacotherapy and behavior therapy for persistent insomnia. *Am J Psychiatry* 2002; 159: 5–11.
6. Wang MY, Wang SY and Tsai PS. Cognitive behavioural therapy for primary insomnia: a systematic review. *J Adv Nurs* 2005; 50: 553–564.
7. Irwin MR, Cole JC and Nicassio PM. Comparative meta-analysis of behavioral interventions for insomnia and their efficacy in middle-aged adults and in older adults 55+ years of age. *Health Psychol* 2006; 25: 3–14.
8. Jacobs GD, Pace-Schott EF, Stickgold R, et al. Cognitive behavior therapy and pharmacotherapy for insomnia: a randomized controlled trial and direct comparison. *Arch Intern Med* 2004; 164: 1888–1896.
9. Manber R, Edinger JD, Gress JL, et al. Cognitive behavioral therapy for insomnia enhances depression outcome in patients with comorbid major depressive disorder and insomnia. *Sleep* 2008; 31: 489–495.
10. Manber R, Carney C, Edinger J, et al. Dissemination of CBTI to the non-sleep specialist: protocol development and training issues. *J Clin Sleep Med* 2012; 8: 209–218.
11. Karlin BE, Trockel M, Taylor CB, et al. National dissemination of cognitive behavioral therapy for insomnia in veterans: therapist and patient-level outcomes. *J Consult Clin Psychol* 2013; 81: 912–917.
12. Trockel M, Karlin BE, Taylor CB, et al. Cognitive behavioral therapy for insomnia with veterans: evaluation of effectiveness and correlates of treatment outcomes. *Behav Res Ther* 2014; 53: 41–46.
13. Clough BA and Casey LM. Technological adjuncts to enhance current psychotherapy practices: a review. *Clin Psychol Rev* 2011; 31: 279–292.
14. Luxton DD, McCann RA, Bush NE, et al. mHealth for mental health: integrating smartphone technology in behavioral healthcare. *Prof Psychol Res Pr* 2011; 42: 505–512.
15. Garrison KA, Pal P, Rojiani R, et al. A randomized controlled trial of smartphone-based mindfulness training for smoking cessation: a study protocol. *BMC Psychiatry*. Epub ahead of print 14 April 2015. DOI: 10.1186/s12888-015-0468-z.
16. Donker T, Petrie K, Proudfoot J, et al. Smartphones for smarter delivery of mental health programs: a systematic review. *J Med Internet Res* 2013; 15: e247.
17. Bucci S, Barrowclough C, Ainsworth J, et al. Using mobile technology to deliver a cognitive behaviour therapy-informed intervention in early psychosis (Actissist): study protocol for a randomised controlled trial. *Trials*. Epub ahead of print 10 September 2015. DOI: 10.1186/s13063-015-0943-3.

18. Monney G, Penzenstadler L, Dupraz O, et al. mHealth app for cannabis users: satisfaction and perceived usefulness. *Front Psychiatry*. Epub ahead of print 27 August 2015. DOI: 10.3389/fpsyt.2015.00120.
19. Tregarthen JP, Lock J and Darcy AM. Development of a smartphone application for eating disorder self-monitoring. *Int J Eat Disord* 2015; 48: 972–982.
20. Erbes CR, Stinson R, Kuhn E, et al. Access, utilization, and interest in mHealth applications among veterans receiving outpatient care for PTSD. *Mil Med* 2014; 179: 1218–1222.
21. Behar J, Roebuck A, Domingos JS, et al. A review of current sleep screening applications for smartphones. *Physiol Meas* 2013; 34: R29–R46.
22. Bhat S, Ferraris A, Gupta D, et al. Is there a clinical role for smartphone sleep apps? Comparison of sleep cycle detection by a smartphone application to polysomnography. *J Clin Sleep Med* 2015; 11: 709–715.
23. Ko PT, Kientz JA, Choe EK, et al. Consumer sleep technologies: a review of the landscape. *J Clin Sleep Med*. Epub ahead of print 22 Jun 2015. DOI: 10.5664/jcsm.5288.
24. Hoffman JE, Taylor KL, Manber R, et al. CBT-I Coach (Version 1.0) (Mobile application software), 2013, http://itunes.apple.com
25. Hoffman JE, Taylor K, Manber R, et al. CBT-I Coach (Version 1.0) (Mobile application software), 2013, https://play.google.com/store
26. Kuhn E, Weiss BJ, Taylor KL, et al. CBT-I Coach: a description and clinician perceptions of a mobile app for cognitive behavioral therapy for insomnia. *J Clin Sleep Med* 2016; 12: 597–606.
27. Babson KA, Ramo DE, Baldini L, et al. Mobile app-delivered cognitive behavioral therapy for insomnia: feasibility and initial efficacy among veterans with cannabis use disorders. *JMIR Res Protoc*. Epub ahead of print 17 July 2015. DOI: 10.2196/resprot.3852.
28. Bastien CH, Vallieres A and Morin CM. Validation of the Insomnia Severity Index as an outcome measure for insomnia research. *Sleep Med* 2001; 2: 297–307.
29. Espie CA. "Stepped care": a health technology solution for delivering cognitive behavioral therapy as a first line insomnia treatment. *Sleep* 2009; 32: 1549–1558.
30. Edinger JD. Is it time to step up to stepped care with our cognitive-behavioral insomnia therapies? *Sleep* 2009; 32: 1539–1541.

Article

Health Informatics Journal

Health Informatics Journal
2018, Vol. 24(1) 14–23

DOI: 10.1177/1460458216652645
journals.sagepub.com/home/jhi
SAGE

Can the British Heart Foundation PocketCPR application improve the performance of chest compressions during bystander resuscitation: A randomised crossover manikin study

Georgette Eaton
Oxford Brookes University, UK

John Renshaw
Coventry University, UK

Pete Gregory
University of Wolverhampton, UK

Tim Kilner
University of Worcester, UK

Abstract

This study aims to determine whether the British Heart Foundation PocketCPR training application can improve the depth and rate of chest compression and therefore be confidently recommended for bystander use. A total of 118 candidates were recruited into a randomised crossover manikin trial. Each candidate performed cardiopulmonary resuscitation for 2 min without instruction or performed chest compressions using the PocketCPR application. Candidates then performed a further 2 min of cardiopulmonary resuscitation within the opposite arm. The number of chest compressions performed improved when PocketCPR was used compared to chest compressions when it was not (44.28% vs 40.57%, $p < 0.001$). The number of chest compressions performed to the required depth was higher in the PocketCPR group (90.86 vs 66.26). The British Heart Foundation PocketCPR application improved the percentage of chest compressions that were performed to the required depth. Despite this, more work is required in order to develop a feedback device that can improve bystander cardiopulmonary resuscitation without creating delay.

Corresponding author:
Georgette Eaton, Oxford Brookes University, Faculty of Health and Life Sciences, Oxford, OX3 0BP, UK.
Email: georgette.eaton@scas.nhs.uk

Keywords
assistive technologies, bystander, healthcare service innovation and information technology, mobile health, PocketCPR

Introduction

The provision of effective bystander cardiopulmonary resuscitation (CPR) in out-of-hospital cardiac arrest (OOHCA) remains unacceptably low,[1–5] despite evidence that suggests that effective bystander CPR is associated with more favourable clinical outcomes and improved survival rates.[6] It is recognised that CPR is frequently inadequate when performed by laypersons, [1,7] with many responders reluctant to perform mouth-to-mouth resuscitation during CPR.[8] To encourage uptake of bystander CPR attempts within the United Kingdom, the British Heart Foundation[8] launched a campaign to promote chest-compression-only CPR, with multiple studies supporting this approach.[9–12] As part of their campaign, the British Heart Foundation produced a smartphone PocketCPR training application to provide real-time feedback on the depth of chest compressions performed during CPR and provide metronomic feedback to ensure accurate external chest compression rate. The importance of adequate chest compressions' depth and accurate rate of compressions were both reaffirmed within current resuscitation guidelines,[13] with suboptimal compression rates associated with poor return of spontaneous circulation.[14] Nevertheless, the performance of both compression depth and compression rate by bystanders is shown to be suboptimal.[15,16]

To improve the performance of chest compressions, feedback systems have been used successfully in training to improve the overall quality of layperson CPR[3,4,6,9,17] and maintain skill acquisition and retention,[17,18] although there has been insufficient evidence to validate these applications for use in practice. This study endeavours to determine whether the British Heart Foundation PocketCPR feedback application would improve chest compression performance during bystander resuscitation.

Materials and methods

Objectives

The aim of this randomised crossover manikin study was to investigate whether using the British Heart Foundation PocketCPR training application would improve the performance of chest compressions against current resuscitation guidelines when used by laypersons with no recent CPR training.

We hypothesised that the British Heart Foundation PocketCPR training application would increase the proportion of chest compressions performed at the recommended depth of 50–60 mm and improve the rate of chest compressions per minute, compared to no application.

Participants and randomisation

Participants were recruited from a university campus on an opportunistic basis. All participants were required to be aged 18 years or over, not be a healthcare professional and not having attended a CPR training course in the last 6 months. This last point was pertinent, since it is acknowledged in the literature that skills and knowledge relating to bystander CPR decay rapidly after initial training.[19–21]

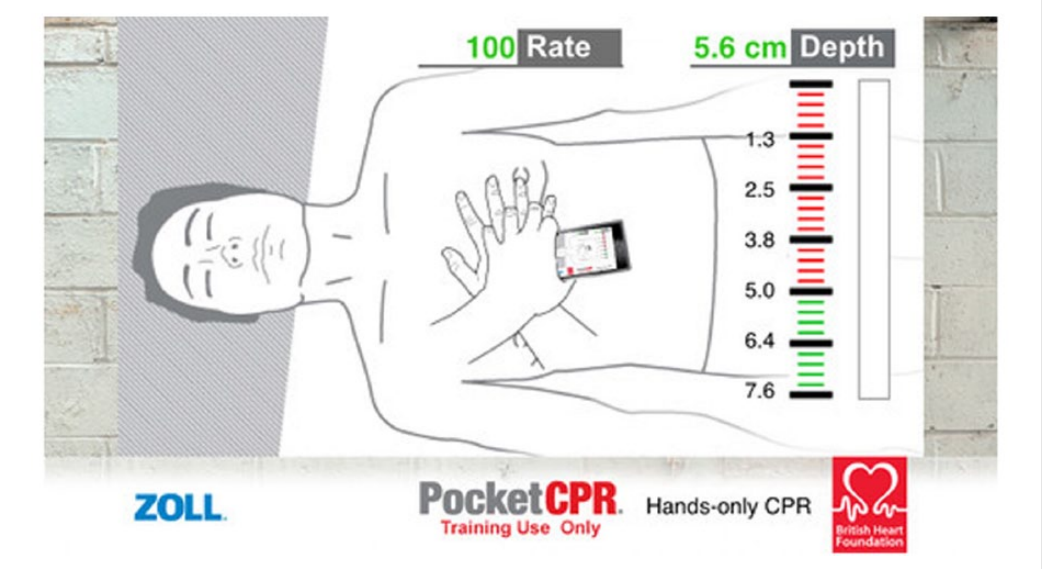


Figure 1. Screen print of PocketCPR feedback screen.

Volunteers were provided with a participant information sheet and an opportunity to ask questions of the researchers before being asked to give written consent.

Methodology

In this randomised crossover study, each participant was asked to attempt a 2-min period of CPR on a Laerdal Resuscitation manikin (Resusci Anne Skills Station, Laerdal Medical Limited, Orpington, UK) with and without the PocketCPR application in accordance with a pre-randomised order. Candidates were not required to have previous experience or ownership of a smartphone device to take part and were provided with information as to how to hold the iPod and activate the PocketCPR software. Randomising the participants reduced the risk that the participant may perform better in the second arm of the study where they had previously used PocketCPR. The randomisation order was generated using the statistical software package PASW (version 17.0.2; SPSS Inc, Chicago, IL, USA) and ensured that 50 per cent of the participants performed first with the PocketCPR application and 50 per cent without. The PocketCPR application was used on an iPod Touch 2009 device.

Participants were required to rest for 2 min between each CPR scenario to ensure that operator fatigue did not adversely affect the second attempt. The instructions given to participants were limited to information on how to hold the iPod and how to activate the British Heart Foundation PocketCPR software.

The software gives visual feedback in the form of a bar on the display indicating current compression depth with a green colour marking the ideal interval and verbal feedback prompts (including 'press harder', 'press faster', 'press slower' and 'good depth') (Figure 1). Additionally, the device has an integrated metronome which ticks at a rate of 100/min, signalling the correct compression rate. There was no feedback on ventilations since this application is designed to support chest-compression-only CPR. When performing CPR without the device, the participant received no verbal or visual feedback and no metronome guidance.

Data collection

Performance measurements derived from the manikin software were recorded onto a connected laptop. These were compression rate, compression depth, hand position for performing chest compressions, tidal volume of ventilation attempt and time off the chest once CPR had been started. Additional observations (including time to start CPR) were recorded manually by the researchers. Manual observations included any ventilation attempt and rate which did not register in the manikin due to an occluded airway.

Statistical analysis

Sample size was calculated based on the primary outcome measure of the adequate depth of compression. Previous research has reported that 42 per cent of trained prehospital providers delivered chest compressions with a mean depth of 50–60 mm during 1-min simulated cardiac arrest scenario, compared with 39 per cent listening to a musical prompt (correlation coefficient (phi) = 0.44051).[22] In order to detect 15 per cent (from 42% to 57%) increase in the proportion of laypersons delivering compressions at the recommended depth with a power of 0.85 and an alpha of 0.05, it was estimated that 108 subjects were required (sample size for paired cohort study calculated using StatsDirect, version 2.7.8; StatsDirectLtd, Altrincham, UK).

Analysis compared the difference in performance of chest compressions with and without the British Heart Association PocketCPR application. The primary outcome measure of mean compression depth was analysed alongside secondary outcome measures of mean total compressions in 2 min of simulated CPR, mean compression rate, mean total correct compressions and correct hand position.

The quality of chest compressions was measured with Resusci Anne Skills Station (Laerdal Medical Limited, Orpington, UK). IBM SPSS Statistics version 22 software package was used to calculate descriptive statistics, p values, 95 per cent confidence intervals (CIs) and Wilcoxon's rank-sum tests for two related samples. A significance level of $p < 0.05$ was considered to be statistically significant.

Ethical considerations

This randomised control trial received ethical approval from the Coventry University Ethics Committee (P4090).

Results

Flow and baseline characteristics

A total of 118 subjects were recruited to the study and were included in the analysis. The sample size of 108 subjects determined by the power sample size calculation was satisfied. Baseline characteristics are shown in Table 1.

Primary outcome

When using the PocketCPR application, 44.28 per cent of the total number of compressions were measured to be the correct depth compared with 40.57 per cent of mean total compressions when the PocketCPR application was not used ($p < 0.001$). The actual number of correct depth compressions

Table 1. Demographic table of participants (n = 120).

Gender	Total
Male	58
Female	62
Age group (years)	Total
18–25	68
26–33	11
34–41	8
41–48	22
49+	10

Table 2. Results.

Parameter	With app. mean %	95% CI	Without app. mean %	95% CI	p
Predicted maximum compressions	200–240		200–240		
Total compressions	205.19	199.11–211.24	163.25	151.24–175.25	0.000
Compression rate	106.87	104.87–108.88	105.37	100.58–110.17	0.858
Correct compressions	30.67 (14.94)	20.38–40.96	20.5 (12.55)	13.03–27.97	0.085
Adequate depth	90.86 (44.28)	75.74–105.99	66.24 (40.57)	52.5–79.97	0.001
Insufficient depth	114.32 (55.71)	99.32–129.33	97.01 (59.42)	81.19–112.82	0.006
Low-hand position	48.1 (23.44)	34.54–61.66	44.97 (27.54)	32.68–57.26	0.970
High-hand position	32.52 (15.84)	21.96–43.08	28.29 (17.32)	18.49–38.09	0.351
Right-hand position	11.42 (5.56)	3.97–18.86	4.41 (2.7)	−0.2–9.02	0.194
Left-hand position	28.36 (13.82)	17.6–39.12	23.77 (14.56)	14.21–33.33	0.788

CI: confidence interval.

was also higher in the PocketCPR group than would be anticipated by the percentages (90.86 vs 66.24), as this group performed more compressions in the 2-min period.

Secondary outcomes

Further analysis was performed on the secondary outcome measures, and the results are reported in Table 2.

The 2010 Resuscitation Council (UK) Guidelines[13] advocate a compression rate of 100–120 compressions per minute; therefore, continuous chest compressions over a 2-min period should result in 200–240 compressions being delivered over the 2-min period. While the mean compression rate when using the PocketCPR was broadly similar to the mean compression rate when not using the application, there was a significant difference in the total number of compressions performed. When using the PocketCPR application, the number of compressions delivered fell within the expected range, while the number in the non-PocketCPR group was lower ($p < 0.000$).

There was no significant difference between mean compression rates, but there was a significant difference between the number of compressions performed during the 2-min time period.

The difference in the mean number of correct compressions when using PocketCPR and without the application was not significant, and the mean number of correct compressions was low in both groups. A total of 14.94 per cent of the mean total compressions (95% CI 20.38–40.96) were correct in terms of rate, depth and hand position when using the application, 12.55 per cent of the mean total compressions (95% CI 13.03–27.97) without the application.

Where compressions were incorrect due to incorrect hand placement, hands were more likely to be too low rather than too high. Where hands were placed away from the midline, subjects were most likely to place their hands further from their body to the left of the midline of the manikin, rather than closer to them to the right of the midline of the manikin.

Discussion

Comment

The PocketCPR training application allowed for a greater depth of compressions during the 2-min resuscitation attempt, which supports the recommendations for CPR in current guidelines.[13] Participants were more likely to reach the recommended resuscitation guidance depth of 50–60 mm for chest compressions when using the PocketCPR application (95% CI 81.19–112.82, $p < 0.006$). With evidence connecting an increased likelihood of return of spontaneous circulation with chest compressions performed to a depth of 50 mm or more,[23,24] the results here suggest that the feedback provided by the PocketCPR application could have real-life applications. Previous manikin-based studies have found that compressions tend towards inadequate depth following only 1 min of CPR due to rescuer fatigue,[25] whereas our results demonstrate that feedback to the layperson allowed chest compressions to be performed to an adequate depth more frequently during a 2-min cycle. This suggests that the PocketCPR application may either ameliorate the effect of rescuer fatigue or help to motivate the rescuer to continue to compress the chest to an adequate depth even when fatigue or a loss of concentration is taking effect.

While the depth of chest compressions improved with the PocketCPR application, it was observed that hand position was frequently reported off-centre (Table 2). This study found that incorrect hand position was often too high on the chest, rather than too low or too far left or right. As a percentage, there was no difference in hand positioning between participants using the PocketCPR device or those without. The use of PocketCPR likely required more dexterity to hold the device between the hands during chest compressions as well as the requirement to visualise the screen to view depth attainment, although our results failed to reach significance. There is insufficient evidence in research to determine if there is any relationship between incorrect hand position and changes in efficiency of CPR.[25] It may also be considered that CPR performed too right or too left is more likely to be effective than chest compressions performed too high and too low due to changes in thoracic pressure as part of the thoracic pump theory.[26]

Despite the high number of incorrect hand positions in both groups, use of PocketCPR did result in a greater number of chest compressions that were performed during the 2-min resuscitation attempt. When using the PocketCPR application, participants achieved the predicted range of 200–240 chest compressions, over the 2-min test period, in accordance with resuscitation guidance of 100–120 chest compressions per minute. Although the number of chest compressions performed using PocketCPR was higher than without, there was a noticeable delay in starting chest compressions while participants navigated the British Heart Foundation PocketCPR training application. While any lack of familiarity with the device was ameliorated by the instructions that were given to the participants, on average, chest compressions were delayed by 37.31 s while candidate's summoned help, confirmed that the patient was not breathing and began chest compressions with device feedback. The delay to begin chest compression is consistent with comparable studies and similar portable feedback devices.[25] Delay to CPR without the device was on average 14.42 s. The

PocketCPR device appears to improve the consistency of chest compressions, as participants still performed a greater number of chest compressions in the required time, albeit with an initial delay. It is unclear whether the delay to start initial chest compressions is countered by the increased number of compressions achieved over 2 min, but the delay in starting could be further exacerbated where the application is not readily available on the home screen of an individual about to perform bystander CPR. A recommendation would be to ensure that the application is pinned to a person's home page on their mobile device in order to minimise any delays in starting chest compressions. To overcome this delay in beginning chest compressions, it is recommended that instruction is more concise, bypassing the approach and navigating straight to resuscitation feedback. Without the application, CPR was often commenced earlier but had lower consistency and longer periods of inactivity due to some participants attempting mouth-to-mouth ventilation. Since bystander resuscitation with periods of inactivity is associated with poorer outcome,[27] it is important to minimise this inactivity. In our study, those performing chest-compression-only CPR using the PocketCPR application had fewer periods of inactivity and more consistency in their compressions. The periods of inactivity serve to explain why the number of compressions in the PocketCPR arm of the study performed significantly more compressions during the 2 min of the study; yet, there was no significant difference in the actual rate of compressions between the two limbs.

During both limbs of the study subjects managed to achieve an average rate of compression that accords with current resuscitation guidelines, which suggest that chest compressions of at least 100 compressions per minute are more effective than slower rates.[28]

Participant recruitment was representative of a normal urban population. Since OOHCA can occur in any environment, the participant demographics are representative of those who may render aid in this situation. During the study, it was noted that most participants were familiar with the device and navigation through the application but were more often unable to effectively follow the instructions as directed. Although we collected information on age, there are many factors that affect the rescuers' ability or willingness to perform bystander CPR. These include (but are not limited to) socioeconomic profile, education,[29] gender and fitness.[30]

Participants in this study were seen to encounter navigation problems while the application was playing due to the touchscreen nature of the device. This resulted in accidental disruption of CPR instruction and restarting the application, which adversely impacted upon the time to perform chest compressions. It would be helpful if the device became locked once the application had been selected and the accelerometers activated by chest compressions. Despite this, the benefit of using the PocketCPR application is that participants only required a smartphone device with the application, rather than additional equipment to secure the device. While other studies have shown improved CPR with the use of smartphones secured in armbands, other studies demonstrate that participants were unable to perform CPR with feedback without the securing mount.[31] While the PocketCPR uses smartphone technology to feedback depth measurement, gripping the device while performing chest compressions is difficult, and the usability is compromised due to the accidental disruption of CPR instruction.

There is a growing body of evidence to suggest that using a smartphone application can improve chest compression depth and rate,[32] which positively affects chance of survival in OOHCA. However, these devices are limited in their usability and complex interfaces, which must be overcome to confidently recommend their use beyond a training application into a real-time feedback device. Smartphone applications may also be useful as prerequiste learning for CPR training.[33]

Although this study considers PocketCPR application to improve chest compression performance, the evolution of portable smart technology is becoming ever more prevalent in prehospital resuscitation, evident by the recent endorsement of the Good Smartphone Activated Medics

(GoodSAM) application from the Resuscitation Council UK,[34] which alerts nearby rescuers to cases of OOHCA. Therefore, there is opportunity to develop applications like PocketCPR to combine rescuer activation with effective CPR feedback until professional help arrives.

Limitations

There were several limitations within this study. First, over-compression of chest compressions was not measurable on the resuscitation manikin, as the physical design of the manikin prevented the participant from over-compressing. Data were collected to consider whether compression depth was insufficient, but it is not known how many compressions may have been too deep. Despite this, there is insufficient evidence to specify an upper limit for chest compression depth, and chest compressions that are too deep may still be effective.[35]

Second, the number of people who have PocketCPR application downloaded onto their portable device or smartphone and accessible during OOHCA incidents may limit the usability of feedback. It is not known how many times the application has been downloaded, but it works on both an Apple and Android platforms, so there is considerable potential for this application be widely available. In the first quarter of 2015 alone, over 74 million iPhones were sold as a stand-alone mobile product.[36] It is not unreasonable to argue that the application should be included as a default application on all devices capable of supporting it.

Finally, as with all simulation and manikin studies, the results of this study cannot measure clinical outcome or survivability but remains a useful proxy measure into the usability of feedback devices for bystanders.

Conclusion

Overall, the standard of bystander resuscitation within this study was poor and chest compressions were still frequently performed at insufficient depth, with incorrect hand positioning and with prolonged periods of inactivity. The PocketCPR application improved the percentage of chest compressions that were performed to the correct depth during bystander compression-only CPR. A greater number of chest compressions were also performed with the application during the 2-min time period when compared to standard basic life support attempts, where compressions were often too shallow and with too few external chest compressions performed. Although use of the application improved CPR performance when compared to no application, CPR performance remained suboptimal. More work is needed to develop an application that can instruct bystanders to perform effective chest-compression-only CPR without delay.

Acknowledgements

The authors would like to thank Malcolm Woollard for his support in the early stages of this study as well as Coventry University for their facilitation.

Declaration of conflicting interests

The author(s) declared no potential conflicts of interest with respect to the research, authorship and/or publication of this article.

Funding

The author(s) received no financial support for the research, authorship and/or publication of this article.

References

1. Spooner B, Fallaha F, Kocierz L, et al. An evaluation of objective feedback in basic life support (BLS) training. *Resuscitation* 2007; 73: 417–424.
2. Iwami T, Kawamura T, Hiraide A, et al. Effectiveness of bystander-initiated cardiac-only resuscitation for patients with out-of-hospital cardiac arrest. *Circulation* 2007; 116: 2900–2907.
3. Sayre R, Berg A, Cave M, et al. Hands-only (compression-only) cardiopulmonary resuscitation: a call to action for bystander response to adults who experience out-of-hospital sudden cardiac arrest: a science advisory for the public from the American Heart Association Emergency Cardiovascular Care Committee. *Circulation* 2008; 117: 2162–2167.
4. Drager K. Improving patient outcomes with compression only CPR: will bystander CPR rates improve? *J Emerg Nurs* 2012; 38: 243–248.
5. Takei Y, Nishi T, Matsubara H, et al. Factors associated with quality of bystander CPR: the presence of multiple rescuers and bystander-initiated CPR without instruction. *Resuscitation* 2014; 85: 492–498.
6. Zanner R, Wilhelm D, Feussner H, et al. Evaluation of M-AID, a first aid application for mobile phones. *Resuscitation* 2007; 74: 487–494.
7. Aufderheide P, Pirrallo G, Yannopoulos D, et al. Incomplete chest wall decompression: a clinical evaluation of CPR performance by trained laypersons and an assessment of alternative manual chest compression-decompression techniques. *Resuscitation* 2006; 71: 341–351.
8. British Heart Foundation. Hands only CPR. *British Heart Foundation*, http://www.bhf.org.uk/heart-health/life-saving-skills/hands-only-cpr-faqs.aspx (2013, accessed 30 March 2015).
9. Urban J, Thode H and Stapleton E. Current knowledge of and willingness to perform Hands-Only™ CPR in laypersons. *Resuscitation* 2013; 84: 1574–1578.
10. Waalewijn RA, Tijssen JGP and Koster RW. Bystander initiated actions in out-of-hospital cardiopulmonary resuscitation: results from the Amsterdam Resuscitation Study (ARRESUST). *Resuscitation* 2001; 50: 273–279.
11. Bobrow BJ, Spaite DW, Berg RA, et al. Chest compression–only CPR by lay rescuers and survival from out-of-hospital cardiac arrest. *JAMA* 2010; 304: 1447–1454.
12. Nagao K, Kikushima K, Sakamoto T, et al. Cardiopulmonary resuscitation by bystanders with chest compression only (SOS-KANTO): an observational study. *Lancet* 2007; 369: 920
13. Nolan J, Soar J, Zideman D, et al. European Resuscitation Council Guidelines for Resuscitation 2010 Section 1. Executive summary. *Resuscitation* 2015; 81: 1219–1276.
14. Abella B, Sandbo N, Vassilatos P, et al. Chest compression rate during cardiopulmonary resuscitation are suboptimal: a prospective study during in-hospital cardiac arrest. *Circulation* 2005; 111: 428–438.
15. Nishiyama C, Iwami T, Kawamura T, et al. Effectiveness of simplified chest compression-only CPR training for the general public: a randomized controlled trial. *Resuscitation* 2008; 79: 90–96.
16. Birkenes T, Myklebust H, Neset A, et al. Quality of CPR performed by trained bystanders with optimized pre-arrival instructions. *Resuscitation* 2014; 85: 124–130.
17. Yeung J, Meeks R, Edelson D, et al. The use of CPR feedback/prompt devices during training and CPR performance: a systematic review. *Resuscitation* 2009; 80: 743–751.
18. Low D, Clark N, Soar J, et al. A randomised controlled trial to determine if the use of iResus© application on a smart phone improves the performance of an advances life support provider in a simulated medical emergency. *Anaesthesia* 2011; 66: 255–262.
19. Woollard M, Whitfeild R, Smith A, et al. Skill acquisition and retention in automated external defibrillator (AED) use and CPR by lay responders: a prospective study. *Resuscitation* 2004; 60: 17–28.
20. Creutzfeldt J, Hedman L, Medin C, et al. Retention of knowledge after repeated virtual world CPR training in high school students. *Stud Health Technol Inform* 2009; 142: 59–61.
21. Isbye D, Meyhoff C, Lippert F, et al. Skill retention in adults and children 3 months after basic life support training using a simple personal resuscitation manikin. *Resuscitation* 2007; 74: 296–302.
22. Woollard M, Poposki J, McWhinnie B, et al. Achy breaky makey wakey heart? A randomised cross-over trial. *Emerg Med J*, http://emj.bmj.com/content/early/2011/10/19/emermed-2011-200187.full (2011, accessed 30 March 2015).

23. Rajab T, Pozner C, Conrad C, et al. Technique for chest compressions in adult CPR. *World J Emerg Surg* 2011: 41, http://www.ncbi.nlm.nih.gov/pmc/articles/PMC3261806/pdf/1749-7922-6-41.pdf (accessed 7 June 2015).
24. Hightower D, Thomas S, Stone C, et al. Decay in quality of closed-chest compressions over time. *Ann Emerg Med* 1995; 26: 300–303.
25. Zapletal B, Greif R, Stumpf D, et al. Comparing three CPR feedback devices and standard BLS in a single rescuer scenario: a randomised simulation study. *Resuscitation* 2014; 85: 560–566.
26. Chul Cha K, Jun Kim Y, Jin Shin H, et al. Optimal position for external chest compression during cardiopulmonary resuscitation: an analysis based on chest CT in patients resuscitated from cardiac arrest. *Emerg Med J* 2013; 30: 615–619.
27. Eftestøl T, Sunde K and Steen P. Effect of interrupting precordial compressions on calculated probability of defibrillation success during out-of-hospital cardiac arrest. *Circulation* 2002; 105: 2270–2273.
28. Nolan J, Perkins G and Soar J. Chest compression rate: where is the sweet spot? *Circulation* 2012; 125: 2968–2970.
29. Papalexopoulou K, Chalkias A, Dontas I, et al. Education and age affect skill acquisition and retention in lay rescuers after a European Resuscitation Council CPR/AED course. *Heart Lung* 2014; 43: 66–71.
30. Reddy K, Murray B, Rudy S, et al. Abstract 224: effective chest compressions are related to gender and body mass index. *Circulation* 2011; 124: A224
31. Semeraro F, Taggi F, Tammaro G, et al. iCPR: a new application of high-quality cardiopulmonary resuscitation training. *Resuscitation* 2010; 82: 436–441.
32. Sakai T, Kitamura T, Nishiyama C, et al. Cardiopulmonary resuscitation support application on a smartphone – randomized controlled trial. *Circ J* 2015; 79: 1052–1057.
33. Park S. Comparison of chest compression quality between the modified chest compression method with the use of smartphone application and the standardized traditional chest compression method during CPR. *Technol Health Care* 2014; 22: 351–358.
34. Resuscitation Council UK. The RC (UK) endorses the GoodSAM app, https://www.resus.org.uk/statements/the-rc-uk-endorses-the-goodsam-app/ (2016, accessed 7 March 2016).
35. Stiell I, Brown S, Christenson J, et al. What is the role of chest compression depth during out-of-hospital cardiac arrest resuscitation? *Crit Care Med* 2012; 40: 1192–1198.
36. Apple inc. Apple reports record first quarter results, *Press Releases*, January, http://www.apple.com/uk/pr/library/2015/01/27Apple-Reports-Record-First-Quarter-Results.html (2015, accessed 25 June 2015).

Health Informatics Journal

Article

Health Informatics Journal
2018, Vol. 24(1) 24–42

DOI: 10.1177/1460458216656471
journals.sagepub.com/home/jhi
SAGE

Detecting hospital-acquired infections: A document classification approach using support vector machines and gradient tree boosting

Claudia Ehrentraut
Stockholm University, Sweden

Markus Ekholm
KTH Royal Institute of Technology, Sweden

Hideyuki Tanushi
Stockholm University, Sweden

Jörg Tiedemann
University of Helsinki, Finland

Hercules Dalianis
Stockholm University, Sweden

Abstract
Hospital-acquired infections pose a significant risk to patient health, while their surveillance is an additional workload for hospital staff. Our overall aim is to build a surveillance system that reliably detects all patient records that potentially include hospital-acquired infections. This is to reduce the burden of having the hospital staff manually check patient records. This study focuses on the application of text classification using support vector machines and gradient tree boosting to the problem. Support vector machines and gradient tree boosting have never been applied to the problem of detecting hospital-acquired infections in Swedish patient records, and according to our experiments, they lead to encouraging results. The best result is yielded by gradient tree boosting, at 93.7 percent recall, 79.7 percent precision and 85.7 percent F1 score when using stemming. We can show that simple preprocessing techniques and parameter tuning can lead to high recall (which we aim for in screening patient records) with appropriate precision for this task.

Corresponding author:
Claudia Ehrentraut, c/o Dalianis, DSV/Stockholm University, P.O. Box 7003, 164 07 Kista, Sweden.
Email: ehrentraut.claudia@gmail.com

Keywords
clinical decision-making, databases and data mining, ehealth, electronic health records, secondary care

Introduction

Patient security in hospitals is crucial. Various risk factors for patients can be found within clinical settings, including hospital-acquired infections (HAIs). HAI is defined as

> [a]n infection occurring in a patient in a hospital or other healthcare facility in whom the infection was not present or incubating at the time of admission. This includes infections acquired in the hospital but appearing after discharge, and also occupational infections among staff of the facility.[1]

HAIs may be caused by medical procedures, for instance, during the implantation of contaminated urinary tract catheters. HAI might also develop in wounds after surgery or occur when micro-organisms spread from person to person, such as during winter vomiting diseases. HAIs pose a public health problem worldwide. A survey conducted under the patronage of World Health Organization (WHO) in 2002 found that for 55 hospitals in 14 countries, an average of 8.5 percent of all hospital patients suffer from HAI.[1]

Many attempts have been made to confine HAIs, for example, better hygiene or manual surveillance performed by infection control professionals, constituting an additional workload for hospital medical staff and hospital management. Nevertheless, the presence of HAIs remains unvaried in modern health facilities. Hospital Information Systems, which are standard in most health facilities today, in combination with the increasing amount of digital data, has pioneered the way for automatic surveillance systems. In the course of this development, research that focuses on the automatic detection of HAI has emerged throughout the past years. The exact approaches vary, ranging from numerous attempts that implement rule-based systems to fewer machine learning–based approaches.

Our study is of an experimental nature and focuses on applying machine-learning techniques to the problem of detecting HAIs. For our task, two well-known learning algorithms, support vector machines (SVMs) and gradient tree boosting (GTB), were applied to the data. The data used in this study comprise patient records provided by Karolinska University Hospital.

The focus of our study lies on the recall values obtained using different classifiers. We aim at approaching 100 percent recall with the highest precision possible, which is a reasonable overall performance in terms of F_1. As presented in the literature,[2] we obtained encouraging results when applying Naive Bayes, SVM and a C4.5 Decision Tree to the problem in an initial approach. Therefore, SVM, in particular, revealed its potential application for our task, as it tendentiously yielded the best results. Thus, we decided to apply SVM once again, this time with tuned parameters. We further applied GTB since it has good classification abilities and interesting data-mining capabilities.[3] The data-mining capability of interest is the ability to interpret the trained classifier. It makes it possible to get a measurement of how important each feature is. This is of interest since it enables us to assess if the features used by the classifier are plausible indicators of HAI. In combination with each of the classifiers, we applied different data preprocessing and feature selection methods, namely, term frequency (TF), lemmatization, stemming, stop word removal, infection-specific terms, term frequency–inverse document frequency (TF-IDF), a combination of lemmatization, respectively, stemming, stop word removal and TF-IDF. The study focused on answering the question regarding whether or not any preprocessing method or parameter tuning would help to increase performance.

Algorithms with high recall are especially suitable for the screening of infections.[4] Thus, this study is an important step toward implementing a system that is expected to constantly screen patient records and determine whether they contain HAI. Automatic HAI screening is especially valuable for medical staff and hospital management, since it would significantly reduce the burden of manually checking patient records for HAI, which is a time-consuming task even for highly trained experts.[5] Instead of analyzing all records, the hospital staff would only have to check those patient records that the system preselected as containing HAI.

Related work

During the past decade, multiple studies have utilized machine learning in the medical domain. See Claster et al.[6] for an overview of some of the more recent papers. The following section presents recent studies that adapt machine-learning approaches to the problem of detecting HAI.

Researchers have aimed at developing a monitoring system that predicts potential HAIs.[7] In that particular study, six classifiers were applied to the problem: Alternating Decision Tree (ADTree), C4.5, ID3, RNA, Decision Tables and nearest neighbor with generalization (NNge). Their data comprise 1520 patient records from the intensive care unit (ICU) of the University Hospital of Oron, Algeria. From this, 17 features were derived, some of which are sex of patient, age of patient, reason for the hospitalization or catheter. They solely measure accuracy, obtaining the highest one of 100 percent with the NNge classifier.

In a study conducted in Taiwan,[8] linear regression (LR) and artificial neural networks (ANNs) were used to predict HAI. The system used structured data. A total of 16 features were extracted from patient records, ranging from demographic, procedural and therapeutic features to features concerning the general health status of the patient. ANNs are trained using back-propagation and conjugate gradient descent. Evaluation of the system was done using an internal test set from the same hospital, as well as an external test set from a different hospital. For the internal test set of 461 hospitalizations, the best result was produced using the ANN approach, reaching a recall of 96.64 percent and a specificity of 85.96 percent. For the external test set consisting of 2500 hospitalizations from different hospitals, LR gave the best result with 82.76 percent recall and 80.90 percent specificity.

In a series of papers,[9–12] researchers around Gilles Cohen addressed the task of monitoring and detecting HAI using data from the University Hospital of Geneva, Switzerland. Their focus lied on the class imbalance, a problem that can be observed in many real-world classifications, especially in the medical domain. They used data from 683 patients, out of which 11 percent were positive cases (contracted HAI) and 89 percent were negative (did not contract HAI). From these records, the researchers collected features, such as demographic characteristics, admission date or admission diagnosis, and applied various techniques in order to detect patients with HAI.

In another study,[9] the researchers tested (1) random and agglomerative-hierarchical-clustering (AHC) oversampling, (2) K-means subsampling and random subsampling and (3) combined AHC oversampling and K-Means subsampling. They compared them using five different classifiers: IB1, Naive Bayes, C4.5, AdaBoost and a symmetrical-margin SVM. They obtained a recall ranging from 49 percent (IB1) to 87 percent (NB) for the five different classifiers when applying combined AHC oversampling and K-Means subsampling. Specificity ranged from 74 percent (NB) to 86 percent (IB1).

In yet another study,[10] the researchers compared a symmetrical SVM against an asymmetrical one. The experiments showed the inadequacy of the symmetrical SVM when dealing with a skewed class distribution. They obtained the highest recall, at 92 percent, with a specificity of 72.2 percent when using the asymmetrical SVM.

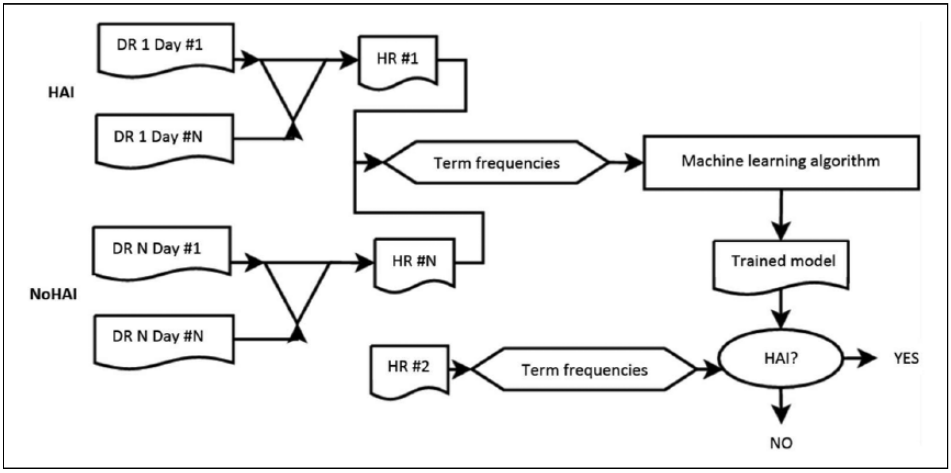


Figure 1. A high-level flow chart describing this study's text-classification approach for automatically detecting HAI. DR stands for daily patient record. In this study, a patient's DR comprises data from four modules. All DRs of a patient together amount to the patient's HR.

In the follow-up paper,[11] the researchers applied one-class SVMs to the problem. This adaption of SVM can be trained to distinguish two classes by ignoring one of the two classes and learning from one class only. Their best results yielded a recall of 92.6 percent at the cost of a very low specificity of 43.73 percent.

In a study from 2006,[12] the researchers compared the resampling strategy that had yielded the best results in a prior study,[9] that is, they combined AHC oversampling and K-means subsampling, to the asymmetrical soft-margin SVM, which had been proven to be suitable for an imbalanced data, as shown in an earlier study.[10] The asymmetrical soft-margin SVM obtained a recall of 92 percent and a specificity of 72 percent, thus clearly outperforming their resampling method that obtained the highest recall at 87 percent with a 74 percent specificity for Naive Bayes.

In two additional studies,[13,14] researchers presented results from a retrospective analysis of data that were collected during the 2006 HAI prevalence survey at the University Hospital of Geneva. The objective of their study, which encompassed both papers, was to define the minimal set of features needed for automated case reporting of HAI. Their dataset comprised 1384 cases, with 166 positive cases (11.99%) and 1218 negative cases (88.01%). The data contained four categories of interest: demographic information, admission diagnosis, patient information on the study date and 6 days before and information related to the infection. They used information gain and SVM recursive feature elimination, combined with chi-squared filtering, to select the most important features. They built two datasets: S1, which contained the most significant features retained by both feature selection algorithms; and S2, which also contained the most important features, but the features that were not well-documented in the patient record were removed. They then applied Fisher's Linear Discriminant for classification. As a result, they obtained 65.37 percent recall and 41.5 percent precision for S1 and 82.56 percent recall and 43.54 percent precision for S2.

Method

The method used in text classification using machine learning, a high-level flow chart, can be seen in Figure 1. An explanation of each part of the flow chart is given in the sections below.

Table 1. The characteristics of the HRs used in our study.

	HAI	NoHAI	Total
Number of HRs	128	85	213
Length of hospitalization in days	2–144	3–93	2–144
Total number of tokens	22,528,102	2,598,036	25,126,138

HR: hospitalization record; HAI: hospital-acquired infection.

Data

The dataset (This research has been approved by the Regional Ethical Review Board in Stockholm (Etikprövningsnämnden i Stockholm), permission number 2012/1838-31/3.) encompasses data from the electronic health records (EHRs) of 120 inpatients at a major university hospital in Sweden and was collected during a Point Prevalence Survey (PPS) (In Sweden, PPSs are performed twice a year to estimate the occurrence of HAI by counting existing cases of HAI at one specific time) in spring 2012. Not all information stored in the patients' EHRs was considered valuable by the physicians for detecting HAI. Thus, a subset of information from the EHRs was retrieved: *Journalanteckning* (Engl.: record notes), *Läkemedelsmodul* (Engl.: drug module), *Mikrobiologiska Svar* (Engl.: microbiological result) and *Kroppstemperatur* (Engl.: body temperature). The information extracted from these modules consists of structured and unstructured data. Structured data refer to data that are stored in predefined fields, such as *International Classification of Diseases*–10th Revision (ICD-10) diagnosis codes, medication or body temperature. Unstructured data refer to textual notes written by physicians, such as daily notes or microbiological results.

For each of the 120 patients, information from all four modules was extracted for the patient's entire hospitalization. The physicians defined one hospitalization as the stay of a patient at a health facility for one care process. If the patient is discharged from one department of the hospital and admitted to another within 24 h, this was regarded as the same hospitalization. Moreover, any noted event occurring within 24 h after discharge was included in the hospitalization. From this point on, we will refer to the file that contains the data of a patient's entire hospitalization as the hospitalization record (HR). Since some of the 120 patients were hospitalized multiple times during the 5-month period of records we received, our dataset comprises 213 HRs. Hospitalizations of less than 48 h are not represented in the final dataset as they were considered to carry too little information. (This time frame is based on international definitions of HAI and the incubation period of infections and is estimated to be less than 48 h for a multitude of disorders.[15]) Table 1 depicts the characteristics of the HRs that were used as input for the classifiers.

All 120 patients had experienced HAI according to the PPS results. We had access to a 5-month period of records. As a result, the physicians in this study, unlike the physicians who carried out the PPSs, obtained information on how the health status of the patient progressed and which assessment he or she received during the time after the PPSs had been conducted. The physicians in this study could therefore give a more accurate answer on whether or not HAI occurred. Only 128 of 213 HRs contained HAI diagnoses (positive examples). Those records represent the HAI class. According to the physician's assessment, the remaining 85 HRs contained no HAI diagnoses (negative examples), thus representing the NoHAI class. The dataset was not balanced, but instead, it was skewed toward the positive class. We only used the class containing HAI for prediction and not the class without HAI.

Machine learning

There are a large number of different learning algorithms and classifier models that could be applied in our classification task. We decided to apply SVMs and GTB to the problem. Instead of exhaustively testing different learning strategies, we focused on a problem that is common for all supervised learning techniques—feature selection and the optimization of parameters. The authors in previous studies[16,17] stated that preprocessing, feature selection and parameter tuning have a large impact on performance—more than the actual choice of the classification model. For a more detailed description of GTB, see Hastie et al.,[18] and for SVM, see Dalal and Zaveri[16] and Noble.[19] The two classifiers are part of the scikit-learn environment (available via http://scikit-learn.org).

SVM

SVMs use the concept of representing the documents that are to be classified as points in a high-dimensional space and finding the hyperplane that separates them. This concept, in fact, is not unique to SVM. However, the difference between SVM and other classifiers using this concept is how the hyperplane is selected. SVM tries to find the hyperplane with the maximum margin, where margin refers to the distance between the hyperplane and the nearest data points.[19] Using SVM is, among others, motivated by the statement that SVM is very effective for two-class classification problems.[16]

We used an optimized and non-optimized SVM on our dataset. For the non-optimized SVM classifier, a radial basis function (RBF) kernel with degree = 3, C = 1, epsilon = 0.001 and gamma = 1/1000 was used. Usually, an RBF kernel is preferred unless the number of features is huge. In that case, a linear kernel is appropriate.[20]

GTB

GTB utilizes the power of a forest of weak tree learners to approximate the sought-after classification function. By training a number of tree classifiers on different parts of the training data and then weighting their collective decision, a strong classifier is produced. The weak learner is a learner that may only have slightly better classification abilities than random guessing, but the combined strong learner will be an approximation of the true classification function. Using decision trees as the weak learner has the advantage of being able to handle different data types without conversion. Inherent when using trees with a maximum depth is feature selection, as only the most important features will be used when constructing the trees. Using trees also makes it possible to interpret the trained model by examining which variables are used most commonly to branch in each individual decision tree.[18] We used GTB both with and without parameter optimization. When used without parameter optimization, the default parameters used were v = 0.1, J = 3, M = 100 and subsample = 1.0.

Preprocessing techniques and parameter optimization

According to previous researchers,[17,21] the high-dimensional feature space, that is, the amount of unique terms that occur in the text documents to be classified, marks a major characteristic and difficulty in text classification, making it a non-trivial task for automatic classifiers. It is thus desirable to reduce the dimensionality of the data to be processed by the classifier, in addition to reducing execution time and improving predictive accuracy. In our study, we used well-known preprocessing and filter methods in order to optimize and reduce the feature space. The preprocessing techniques are depicted in Table 2.

Term frequency (TF)

In this method, TF 1000, the 1000 most frequent terms, was chosen based on their TF. TF refers to the simplest weighting scheme, where the weight of a term is equal to the number of times the term occurs in a document.[22,23]

Table 2. Different combinations of applied text-classification techniques and feature selection methods as well as the name chosen for each combination.

Name	Text-classification method	Feature selection method
TF 1000	Data not processed	TF 1000
Lemma	Data lemmatized	TF 1000
Stem	Data stemmed	TF 1000
Stop	Stop words removed from data	TF 1000
IST	Data not processed	Infection-specific terms used
TF-IDF 1000	Data not processed	TF-IDF 1000
LS-TFIDF 1000	Data lemmatized + stop words removed	TF-IDF 1000
SS-TFIDF 1000	Data stemmed + stop words removed	TF-IDF 1000

TF: term frequency; IST: infection-specific terms; TF-IDF: term frequency–inverse document frequency.

Lemmatization and stemming

In machine learning, lemmatization and stemming are the frequently used methods when preprocessing data.[16] In our study, we use the CST lemmatize (http://cst.dk/online/lemmatiser/uk/) in order to perform lemmatization. Lemmatization describes the process of reducing a word to a common base form, normally its dictionary form (lemma). This is achieved by removing inflectional forms and sometimes derivationally related forms of the word, by means of vocabulary usage and morphological analysis, for instance, *am, are, is, be*, or *hospitals, hospital's → hospital*.[22,23] For the Swedish language, which is highly inflectional, lemmatization is more important than it is for English.

We further use stemming, which is a simpler form of lemmatization, where the produced stemmed words do not need to be real words but the minimal set of characters that distinguish the different stemmed words, for example, *hospitals, hospital's → hospit*. We used the Snowball stemmer (http://snowball.tartarus.org/algorithms/swedish/stemmer.html) for the Swedish language. The patient records were lemmatized and stemmed separately, before then being given as input for the classifiers.

Stop word removal

Stop words are terms that are regarded as not conveying any significant semantics to the texts or phrases they appear in and are consequently discarded.[24] The filter was configured to use the Swedish stop list, which is available via Snowball (http://snowball.tartarus.org/algorithms/swedish/stop.txt) and comprises 113 words, such as *och* (Engl.: and), *att* (Engl.: to) or *i* (Engl.: in).

Infection-specific terms

In the course of the Detect-HAI project (Detection of HAIs through language technology project—conducted in collaboration between Karolinska University Hospital and the Department of Computer and System Science (DSV) at Stockholm University during 2012 and 2013; the aim of the project was to ultimately build a system that can automatically detect HAI in Swedish patient records), a terminology database containing infection-specific terms was built using a semi-automatic approach. Infection-specific terms, such as *kateter* (Engl.: catheter), *ultraljud* (Engl.: ultrasound), *operation* (Engl.: surgery) or *feber* (Engl.: fever), are expected to be contained in patient records in case an infection occurs. In order to build the terminology database, the medical experts involved in that project supplied a seed set of about 30 infection-specific terms, which were based on frequent observations in the above-mentioned data and their knowledge about infections. The seed set was then extended by giving each term of it as input for an automatic synonym extractor. The synonym generator used was implemented in-house and was based on random indexing.[25] For each input term, a table holding related terms, which could include synonyms or misspellings, was generated as an output by the synonym generator. One medical expert then manually analyzed all proposed terms with respect to whether or not they could be regarded as applicable infection-specific terms. All relevant terms were added to the terminology database. The final infection-specific term (IST) terminology database comprised a total of 1045 terms. When using the terminology database as a feature reduction technique, we removed all terms from the HRs except for those that occurred in the terminology database. By means of this procedure, the feature space was decreased to 374.

Term frequency–inverse document frequency (TF-IDF)

In a final approach, we assigned a TF-IDF weight to all terms. TF is defined in section "Term frequency (TF)." IDF is, according to previous research,[22,23] a mechanism used in combination with TF to attenuate the effect of words that occur too often in the set of documents, as they could be important in order to discriminate between those. IDF is calculated as follows: $\mathrm{idf}_t = \log N / \mathrm{df}_t$, where N is the number of documents in a collection and df_t is the document frequency of term t, that is, the number of documents in the collection that contain t. TF-IDF for a term is calculated using: $\text{tf-idf}_{t,d} = \mathrm{tf}_{t,d} \times \mathrm{idf}_t$. Thus, TF-IDF for a t is large if t occurs many times within a small number of documents. We reduced the number of features to a maximum of 1000 terms with the highest TF-IDF scores. For more information on TF-IDF and different weighting schemes, see Manning et al.[22] and Van Rijsbergen.[23]

Combination of preprocessing techniques

In an additional preprocessing step, lemmatization, stop word removal and TF-IDF 1000 were combined. This preprocessing step is named LS-TFIDF 1000 in Table 2. The corresponding step with stemming is named SS-TFIDF. (For more examples of different preprocessing and filtering techniques, see Dalal and Zaveri,[16] Yang and Pedersen[21] and Doraisamy et al.[26]).

Parameter optimization

The chosen machine-learning algorithms have a number of parameters that can be fine-tuned to better adapt to the problem and data they are applied on. For SVM using the RBF kernel, there are two main parameters: C and gamma.[20] C controls the number of misclassified examples tolerated in the training set, while the gamma value affects the number of support vectors used.

In the case of GTB, the important parameters are J, v and M.[18] J refers to the number of terminal nodes in each tree and reflects the number of variable interactions that are possible, v is the learning rate and M is the number of trees. It is usually beneficial to use subsampling with GTB. When using subsampling, a random sample of a predefined size is used to train each tree. This reduces the risk of overfitting. We chose to use 0.5 subsampling, as it is a commonly chosen strategy and has been proven to work well for the task. The learning rate v was fixed to 0.01 after some initial experiments.

In order to find good combinations of parameter values, a grid-search was conducted using fivefold cross-validation on the training data in each fold. Using fivefold cross-validation instead of a higher value of K avoids overfitting, given the small amount of available data. The parameters searched for GTB were as follows:

- $J \Diamond \{1, 3, 6, 8\}$;
- $M \in \{25, 50, 100, 200, 500, 1000\}$.

The parameters searched for SVM were as follows:

- Gamma $\in \{1.0, 0.1, 0.01, 0.001, 0.0001, 0.00001, 0.00001, 0.000001\}$;
- C $\in \{1, 1.5, 2, 2.5, 3, 3.5, 4, 4.5, 5, 5.5, 6, 6.5, 7, 7.5, 8, 8.5, 9, 9.5, 10\}$.

Evaluation

10-fold cross-validation

For evaluation, we use stratified 10-fold cross-validation, which is one of the best known and most commonly used evaluation techniques. Cross-validation is especially useful if the dataset is small, such as in our case, as it maximizes the amount of training data.[27]

Statistical tests

When comparing the classifiers' results, statistical testing is necessary in order to verify the significance of the results. In this study, the non-parametric sign test was used. The choice was motivated by the fact that the authors in previous research[27] presented this statistical test as being simple to calculate and yet appropriate when wanting to compare the performance of multiple classifiers on a single domain. Just like the researchers[27] did in their example calculations, the sign test was one-tailed and performed at 5 percent significance level.

Results

Table 3 depicts the results of SVM, GTB and their optimized counterparts, given the different preprocessing and feature selection methods. The best precision, recall and F1 scores for each preprocessing method are highlighted. For both classifiers, we built the models using both classes, that is, the 128 HRs containing HAI and 85 HRs not containing HAI. However, since the focus of this study lies on obtaining high recall for HRs that contain HAI, we only present performance measures of the classifiers for those records. Precision, recall and F1 scores for HRs not containing HAI are thus neither depicted nor analyzed.

When considering the recall score, one needs to take the F1 score into consideration. A baseline majority classifier would classify all instances as HAI, yielding a recall of 100 percent, precision of 60 percent and F1 score of 75 percent. Hence, if the F1 score for a classifier is close to the baseline of 75 percent, the result is not of interest, even if the recall value is high, as the performance is not better than is the baseline majority classifier. This means that we can disregard all of the results with a recall value of 100 percent since these do not have an F1 score larger than 75 percent. The optimized GTB yields the highest recall for all preprocessing techniques, with a maximum recall value of 93.7 percent and an F1 score of 85.7 percent when using stemming as the preprocessing technique.

When comparing the unoptimized SVM with the optimized SVM approach, it becomes clear that it is very important to perform parameter optimization when using SVM; the optimized SVM obtains a higher F1 score than does the unoptimized SVM for all preprocessing techniques. It is also worth noting that the F1 scores of the unoptimized SVM only differ slightly from the baseline. In the case of GTB, however, the results did not differ much when parameter optimization was applied, and the default parameters used produced a result as good as, slightly better or slightly worse than its optimized counterpart.

To statistically verify the classifiers' different performance, we applied the non-parametric sign test, as mentioned earlier when using stemming as preprocessing, since it yielded the highest recall

Table 3. Precision, recall and F1 score (in %) for detecting HAIs using GTB, optimized GTB, SVM and optimized SVM given the different preprocessing methods.

	GTB			GTB optimized			SVM			SVM optimized		
	P	R	F1	P	R	F1	P	R	F1	P	R	F1
TF 1000	83.4	90.6	86.7	79.6	92.2	85.2	76.3	79.8	78.0	80.2	88.1	83.7
Lemma	79.3	87.6	83.0	76.5	92.2	83.1	60.1	100.0	75.1	78.9	88.2	83.1
Stem	82.4	88.3	85.0	79.7	**93.7**	85.7	60.1	100.0	75.1	80.7	89.8	84.8
Stop	79.0	83.6	80.6	79.0	93.0	85.0	76.5	78.3	77.4	83.1	89.8	84.8
IST	79.0	86.0	81.7	76.7	89.1	81.9	73.0	65.1	68.9	72.9	84.5	78.0
TF-IDF 1000	81.7	91.2	86.0	79.5	92.1	84.9	60.1	100.0	75.1	78.1	89.7	82.8
LS-TFIDF 1000	80.2	84.4	81.9	78.9	91.3	84.2	60.1	100.0	75.1	72.7	88.9	79.3
SS-TFIDF 1000	78.6	85.8	81.6	78.8	93.0	85.0	60.1	100.0	75.1	75.3	86.6	79.8

GTB: gradient tree boosting; SVM: support vector machine; TF: term frequency; IST: infection-specific term; TF-IDF: term frequency–inverse document frequency.

In total, the material comprised 213 HRs of which 128 contained HAI giving a baseline precision of 60 percent, recall of 100 percent and F-score of 75 percent.

score, in combination with a high F1 score for GTB, optimized GTB and optimized SVM. The conclusion was that the performance results obtained using the GTB, optimized GTB and optimized SVM, respectively, are not significantly different. However, they are significantly better compared to the unoptimized SVM that does not perform significantly better than the baseline classifier.

When comparing the recall, precision and F1 scores that the optimized GTB and optimized SVM obtained for the different preprocessing methods, it became apparent that the techniques did not generate a significant difference in the results. For the optimized GTB, the obtained recall values ranged from 89.1 percent (GTB-IST) at the lowest to 93.7 percent (optimized GTB-Stem) at the highest, indicating significant improvement from using one or the other technique. Likewise, the precision and F1 score values did not show any significant difference. The same can be stated for the performance results of the optimized SVM. The recall values varied between the minimum recall of 83.7 percent and the maximum recall of 89.8 percent, not differing significantly.

To summarize our observations, GTB obtained the highest recall and F1 score for all preprocessing methods when results too close to the baseline are disregarded. The difference between the best recall value, 93.7 percent (Stem), and the second best, 93.0 percent (SS-TFIDF 1000 or Stop), was only 0.7 percentage points, and thus was not statistically significant. The difference in the third best recall value, 92.2 percent (TF 1000), amounted to 1.5 percentage points. Compared to this spread, the respective F1 scores remained quite close: 85.7 percent (Stem), 85.0 percent (stop word removal/ SS-TFIDF 1000) and 85.2 percent (TF 1000), indicating a comparable overall performance. The highest F1 score, 85.7 percent, was obtained when only stemming was applied. Since we aimed for the highest recall with the highest precision possible, that is, a reasonable overall performance in terms of F1, we concluded that the performance of optimized GTB-Stem came closest to our objective.

Decision features

It is interesting to look at the features upon which the classifiers base their decision. Using GTB, a measure of the relative importance of each feature used can be obtained by examining and scoring features that are most frequently used to branch off in each tree.[18] The way to visualize and interpret the results happens in the form of a relative importance plot: the value of each feature is calculated as feature value = 100×(feature score / max score), giving each feature a value relative to the most important feature.

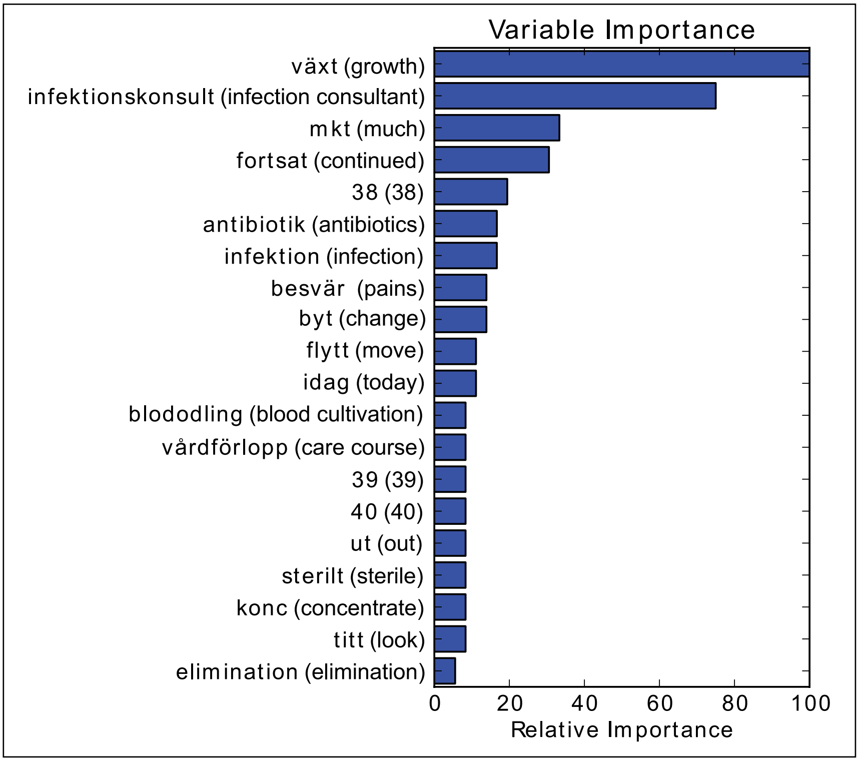


Figure 2. Top 20 feature importances for optimized GTB TF1000+stemming trained on the whole dataset. English translation within parenthesis. Note that since stemming is used the english translation is an approximation as directly translating a stem is not always possible.

Doing this for a GTB classifier trained on the whole stemmed point prevalence measurement (PPM) dataset yielded Figure 2. This can be compared to Figures 3 and 4, which show relative feature importance for unoptimized GTB-Stem and unoptimized GTB using TF1000 without stemming. Based on these figures, some important observations can be made: (1) among the most important features are words that are plausible HAI indicators, such as *växt* (Eng.: growth), *infektionskonsult* (Eng.: infection consultant), *antibiotik* (stemmed Swedish word, Eng.: antibiotics) and *infektion* (Eng.: infection). This is good as it hints that the approach was not building a model randomly. The features should be important indicators, even for larger datasets. (2) We can, however, also observe that *idag* (Eng.: today) and the abbreviation *mkt* (Eng. much), which are considered to be Swedish stop words, were seen as important features. This observation asks for a more thorough analysis of the terminological structure of patient records in order to optimize feature selection. (3) The most important features were independent of parameter optimization and preprocessing: the top two features in all of the figures were *växt* (Eng.: growth) and *infektionskonsult* (Eng.: infection consultant). This observation strengthens the case that the application of different preprocessing techniques may not be very significant. (4) Furthermore, the two most important features were not present in the database of ISTs in section "Infection-specific terms." In other words, there were terms that were either overlooked or deemed to not be indicators of HAI that were, in fact, important. This indicates that feature selection should either be automatic or semiautomatic since there may be important terms that may be left out otherwise.

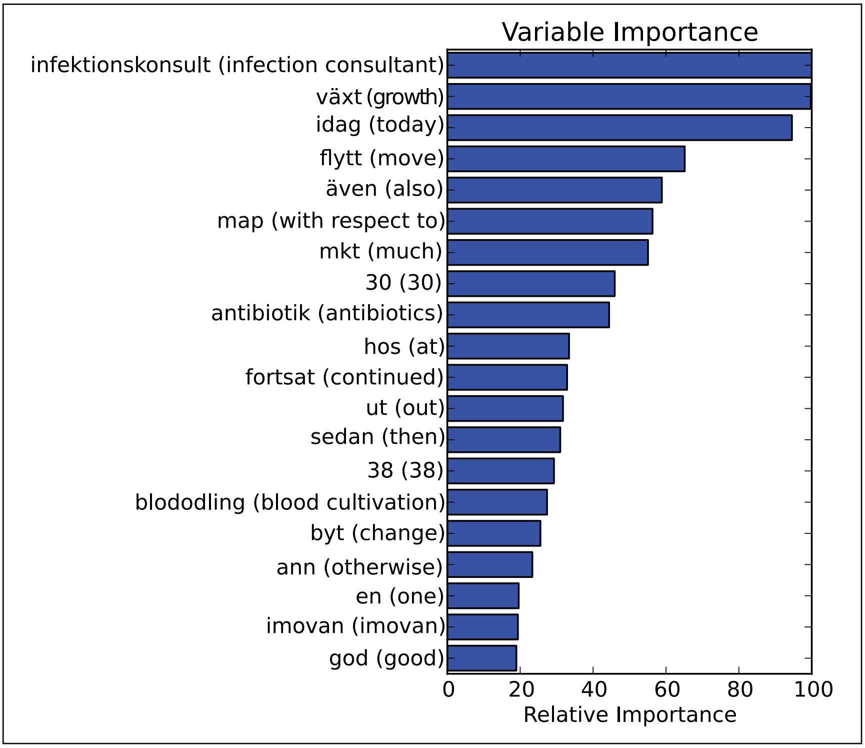


Figure 3. Top 20 feature importances for un-optimized GTB TF1000+stemming trained on the whole dataset. English translation within parenthesis.

Classifier errors

As the optimized GTB-Stem produced the overall best results for our objective, it is interesting to analyze the types of errors the classifier made. To do this, all misclassified examples from each of the 10-folds were examined. It is observable from Table 3 that a recall of 100 percent can be achieved, yet at the cost of very low precision. However, as stated earlier, we emphasized recall (aiming at 100 percent) with the highest precision possible. If we considered the obtained recall of above 90 percent as being sufficiently high, it would be interesting to look at what keeps the precision low. In order to do so, we must evaluate what type of errors were made in the NoHAI class since misclassifications for this class appeared as false positives in the predicted HAI class, keeping the precision score below 80 percent in almost all cases.

As visible in Table 4, the class NoHAI can be divided into two disjoint subclasses: hospitalizations with no infections at all (NoINF) and hospitalizations with community-acquired infections (CAIs), the latter of which we defined as infections that were not acquired in the hospital. Furthermore, some of the hospitalizations were, at the time the PPMs were carried out, considered to contain HAI, but in retrospect did not contain any HAI, but rather, some other type of infection or no infection at all. These are referred to as “HAI suspects.”

If we examine the type of errors made in the NoHAI class, we can make the following observations: 11/14 of all the “suspected HAI” cases were misclassified in comparison to the non-suspects, which were only misclassified in 17/70 of the cases. Furthermore, of all hospitalizations that contained CAI, 12/22 were misclassified, while only 16/62 of all hospitalizations not containing infections were misclassified. Based on this, it seems like it is difficult for the classifier to distinguish

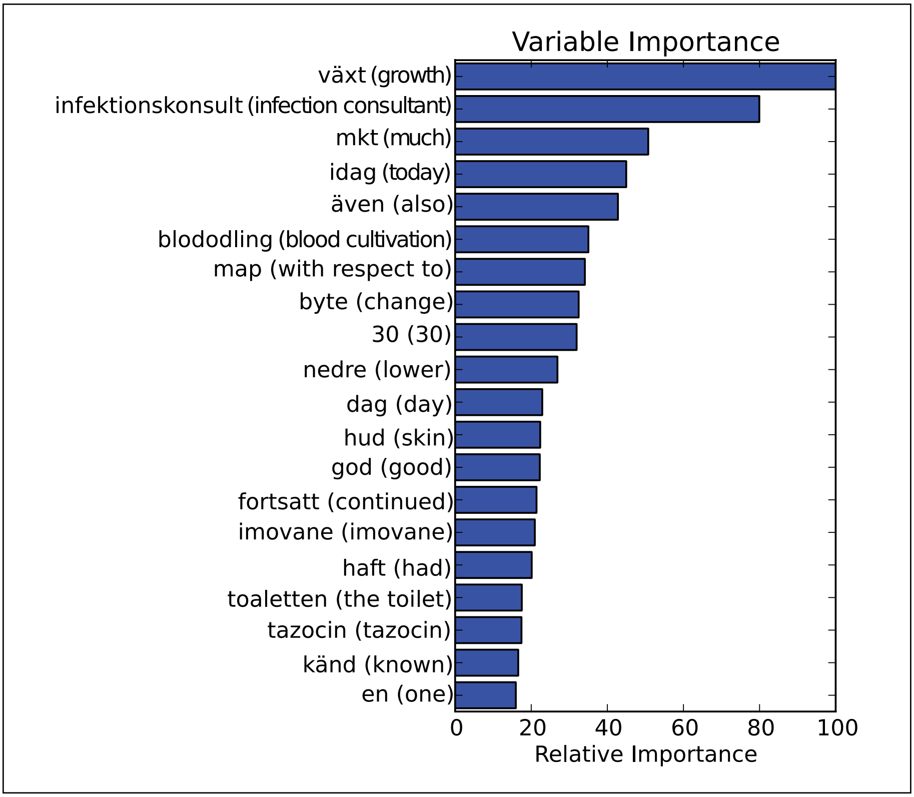


Figure 4. Top 20 feature importances for unoptimized GTB using TF1000 without stemming. English translation within parenthesis.

between a HAI and CAI, and the cases misclassified by the medical staff during the PPM study are indeed hard to classify.

Compared to the fairly high recall, the precision stayed below 80 percent in almost all cases. Error analysis revealed that 42.8 percent of the false positives were patient records that contained a CAI. Another 25 percent of the false positives were "suspected HAI." Handling these false positives in a future approach is crucial for increasing precision. Excluding records containing CAI from classification is one option. Another idea is to, as a first step, train a classifier to differentiate between patient records containing an infection (including HAIs and CAIs) and those not containing an infection. In the second step, HAIs are then detected from the records that were predicted as infections.

Table 5 depicts the classifier errors for each label in class HAI. All hospitalizations that contained ventilator-associated pneumonia (VAP) were classified correctly. Likewise, the classifier performed well for the hospitalizations containing pneumonia and sepsis: only 3 and 2 percent, respectively, of the hospitalizations containing these types of HAI were classified incorrectly. On the other hand, 20 percent of all hospitalizations containing urinary tract infections (UTI) and *Clostridium difficile* were classified into the wrong class. The analysis gives an indication that some types of HAI are easier to classify compared to others.

Yet, the count of certain types, for example, VAP, central venous catheter–related HAI or *Clostridium difficile*, is quite low. It is therefore difficult to say whether or not the error rate would be the same for a larger number of cases. Looking at how the different types of HAIs were classified, it is evident that some types, such as VAP, sepsis and pneumonia, have a lower

Table 4. Classifier errors (optimized GTB-Stem) for the classes HAI and NoHAI, the latter being divided into four disjoint subclasses.

Class structure			Errors	Dataset
HAI			11	128
NoHAI	CAI	Suspected HAI	4	5
		Not suspected HAI	8	18
	NoINF	Suspected HAI	7	9
		Not suspected HAI	9	53
Total			39	213

GTB: gradient tree boosting; HAI: hospital-acquired infection; CAI: community-acquired infection; NoINF: no infections at all.

Table 5. Classifier errors (optimized GTB-Stem) for the different types of HAIs.

Label	Errors	Dataset
Ventilator-associated pneumonia	0	8
Sepsis	1	46
Pneumonia	1	33
Other HAI	1	15
Fungus/virus	1	15
Central venous catheter–related HAI	1	10
Wound infection	2	25
Urinary tract infection	4	20
Clostridium difficile	2	10

GTB: gradient tree boosting.
A hospitalization marked with HAI may have one or more types of HAI. Hence, a misclassified HAI hospitalization may contribute to the number of errors for multiple labels.

error rate compared to, for instance, UTI or *Clostridium difficile*. This observation demands a more thorough analysis of the records and how they have been classified with regard to which type of HAI they contain. This is to ultimately find out whether there are any indications in the records that keep the classification error rate for some types of HAI low, while others remain high.

Discussion

To our knowledge, our project group is the first or only one that has applied machine-learning techniques to Swedish patient records in order to detect HAIs. Compared to our previous approach, which was previously presented,[2] we increased performance while using the same input data, that is, from 89.84 percent recall, 66.9 percent precision and 76.7 percent F1 score when applying SVM-tfidf50 in our previous paper to 93.7 percent recall, 79.7 percent precision and 85.7 percent F1 score when applying optimized GTB-Stem in our present approach. Although the recall values differed by only 3.86 percentage points, we could increase the precision significantly by 12.8 percentage points. This yielded a considerably better F1 score, bringing us an important step toward our aim of approaching a 100 percent recall with the highest precision possible. Moreover, our experiments suggested that we can achieve better results than can some of the approaches presented in section "Related work," even though they are not directly comparable due to different datasets and variant languages used.

Limitations

One limitation of this study was the size of the dataset. The task of manually classifying patient records as to whether or not they contained HAI was difficult and time consuming. Manually analyzing a larger amount of patient records than the one we had and marking them as HAI or NoHAI have therefore not been possible in the amount of time available.

Another limitation was the distribution of positive and false cases in our dataset. The number of records that contain HAI (61%) and NoHAI (39%) in our dataset does not relate to the real-life distribution, which is approximately 10 percent HAI and 90 percent NoHAI. Furthermore, the majority of the NoHAI records were from patients who had a HAI at some point or another (there also exists a HAI record for the same patient). This, as well as the fact that several of the NoHAI records were incorrectly classified as HAI by the medical staff at some point, gives us a dataset where the NoHAI class is harder to distinguish from the HAI class that what we thought would be the case.

Importance of preprocessing methods and choice of classifier

Even though the optimized GTB-Stem yielded the best performance results, it became apparent that the results yielded by a classifier, given the different preprocessing and feature selection techniques, were only marginal. This means that, given our data, it does not make a significant difference whether we choose, for instance, stemming, stop word removal or SS-TFIDF 1000 as a preprocessing technique since the performance results are nearly the same. In our case, it made a bigger difference as to which classifier was used and whether the parameters were tuned. In terms of recall and F1 score, the optimized GTB generally yielded better results than did the (un)optimized SVM. Moreover, the improvement in results obtained using the optimized SVM compared to the unoptimized SVM were clearly visible.

Text classification

A major difference between our approach and the similar work mentioned in Ehrentraut et al.[2] is the fact that we treated all available data—free-text and lab results were treated as a single unstructured text document. This allowed us to apply standard text-classification methods, namely, applying TFs, such as features and standard machine-learning algorithms, to the problem. This has a big advantage compared to methods that rely on structured data: the amount and type of structured data available are different between hospitals and journal systems. The approach does not rely on the availability of such data and it does not rely on the data being available in a certain format. Furthermore, our approach was able to detect HAI indicators that were not known ahead of time, as shown in section Decision Features.

Comparing the text-classification approach with an approach based on structured data requires further research—evaluating and comparing the text-classification approach with a structured data approach using the same dataset. The results of this study showed that in terms of recall, the text-classification approach was proven to produce results that were as good as using structured data (see Table 6). However, the specificity was lower than were the best scores found in the studies using structured data, but this may be due to the characteristics of the dataset used. If the text-classification approach was able to produce the results seen in Table 3 on larger datasets, it might

Table 6. Recall, specificity and precision for optimized GTB-Stem compared with the results found in the "Related work" section.

	Recall	Specificity	Precision
GTB-Stem optimized	93.7	64.1	79.7
[7] ANN Internal	96.64	85.96	–
[7] LR external	82.76	80.90	–
[10] SVM	92.6	43.73	–
[11] SVM	92.0	72.0	–
[11] NB	87	74.0	–
[13] FLD S2	82.56	–	43.54

GTB: gradient tree boosting; ANN: artificial neural network; LR: linear regression; SVM: support vector machine; NB: Naïve Bayes classifiers; FLD: Fisher's linear discriminant.
Note that the evaluation methods and datasets are not the same.

be a good candidate for application in real-world scenarios, with minimal changes to the journal systems used and minimal additional work for the medical staff.

Cost analysis

In a report by the Swedish National Board of Health and Welfare,[28] it is estimated that HAIs prolong the length of a patient's stay in the hospital by an average of 4 days. Given all of the patients suffering from HAI, this is estimated to be about 500,000 extra hospital days per year. With a daily average cost of SEK7.373 (US$860) per day of care, HAIs generate an additional cost of approximately SEK3.7b (US$0.43b) per year except for the labor-intensive cost for carrying out manual PPM twice a year. However, if we can automatize this process, we would save labor, as well as improve the quality of the controls by carrying them out automatically and continuously, *24 h a day, all year long*.[12,29]

Future work

We are well aware of the facts that

- Our dataset is small, containing only 213 instances;
- The distribution of positive and negative cases, that is, 128 HAI and 85 NoHAI instances, does not correlate with the real-life distribution;
- The differences in the results of optimized GTB and optimized SVM are marginal and are not significantly different.

However, the result is significantly better compared to a majority baseline classifiers and unoptimized SVM, and we are convinced that the results reveal the potential for applying text-classification techniques to patient records, including the structured as well as unstructured parts. This is further motivated by the fact that, so far, we have used no particularly elaborate preprocessing and feature reduction methods. Future research will thus have to focus on improving the scores by, for instance, using wrapper techniques for feature reduction that are optimized on a specific learning algorithm and, therefore, yield better results according to previous research.[30]

Moreover, the medical experts involved in this project will manually analyze 292 additional HRs from the rheumatic clinic at Karolinska University Hospital. Thus, we will be able to train the classifiers on about twice as many data as we did for the current project, leading us to expect an improvement in performance. In addition, we aim at training the classifiers on a more realistic dataset, with a real-life distribution of about 10 percent positive and 90 percent negative cases.

Conclusion

This article focuses on applying SVM and GTB to the problem of detecting HAI in digital patient records. By means of applying different preprocessing, as well as feature selection, methods, we tried to increase recall. The results of the machine-learning algorithms were all in all very encouraging. Optimized GTB-Stem came closest to the objective of obtaining high recall with the highest precision possible, that is, yielding a recall of 93.7 percent, precision of 79.7 percent and F1 score of 85.7 percent.

This revealed the applicability of GTB to the task. The increased recall value obtained with the optimized SVM compared to the unoptimized SVM confirmed the assumption that SVM seemed to be suitable for the task and, more importantly, revealed the importance of parameter tuning, leading to significantly better results. Applying stemming yielded high performance results for all three classifiers, yet the difference in the results yielded by the classifiers when other preprocessing techniques (especially stop word removal) are applied were marginal.

Finally, the overall goal will continue to be obtaining high recall (approaching 100%) with the highest precision possible for HRs. This will enable us to implement a system that can screen all HRs and filter out all HRs that contain HAI. This would reduce the workload for hospital staff tremendously as they only need to analyze those HRs that were preselected by the system.

Acknowledgements

We would like to thank our excellent physicians, Maria Kvist and Elda Sparrelid, for their insightful classifications of hospital-acquired infections, as well as Martin Duneld for suggestions regarding preprocessing and computing methods. C.E. did the main work, together with M.E. C.E. wrote the main part of the article and contributed to the data analysis, while M.E. ran the experiments, did the data analysis and contributed to the writing. H.T. extracted and prepared the datasets from the patient record system. J.T. and H.D. gave feedback on the writing and the analysis of the results.

Declaration of conflicting interests

The author(s) received no financial support for the research, authorship, and/or publication of this article.

Funding

The author(s) disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: This work was supported by VINNOVA's (Swedish Governmental Agency for Innovation Systems) project *Innovation mot Infektion* (grant number 2012-01252, 2012) and HIPPA (Hospital Intelligence for Better Patient Security) (grant number 2013-00677, 2013).

References

1. Ducel G, Fabry J and Nicolle L. *Prevention of hospital-acquired infections: a practical guide*. 2nd ed. Geneva: World Health Organization, 2002, p. 1.
2. Ehrentraut C, Tiedemann J, Dalianis H, et al. Detection of hospital acquired infections in sparse and noisy Swedish patient records. In: *Proceedings of the sixth workshop on analytics for noisy unstructured text data (AND 2012)*, Mumbai, India, 9 December 2012, pp. 1–8. New York: ACM.
3. Hastie T, Tibshirani R and Friedman J. *The elements of statistical learning: data mining, inference and prediction*. 2nd ed. New York: Springer, 2008, p. 758.

4. Klompas M and Yokoe DS. Automated surveillance of health care-associated infections. *Clin Infect Dis* 2009; 48(9): 1268–1275.
5. Blacky A, Mandl H, Adlassnig KP, et al. Fully automated surveillance of healthcare-associated infections with MONI-ICU: a breakthrough in clinical infection surveillance. *Appl Clin Inform* 2011; 2(3): 365–372.
6. Claster WB, Shanmuganathan S, Ghotbi N, et al. Text classification for medical informatics: a comparison of models for data mining radiological medical records. *Asia Pac World* 2011; 2(1): 121–137.
7. Benhaddouche D and Benyettou A. Control of nosocomial infections by data mining. *World Appl Program* 2012; 2(4): 216–219.
8. Chang YJ, Yeh ML, Li YC, et al. Predicting hospital-acquired infections by scoring system with simple parameters. *PLoS ONE* 2011; 6(8): e23137.
9. Cohen G, Hilario M, Sax H, et al. Data imbalance in surveillance of nosocomial infections. In: *Proceedings of the medical data analysis: 4th international symposium (ISMDA 2003)* (ed Perner P, Brause R and Holzhütter HG), Berlin, 9–10 October 2003, pp. 109–117. Berlin, Heidelberg: Springer.
10. Cohen G, Hilario M, Hugonnet S, et al. Asymmetrical margin approach to surveillance of nosocomial infections using support vector classification. In: *Proceedings of the intelligent data analysis in medicine and pharmacology (IDAMAP 2003)*, Protaras, 19–22 October 2003, pp. 1–13.
11. Cohen G, Hilario M, Sax H, et al. An application of one-class support vector machine to nosocomial infection detection. In: *Proceedings of the 11th world congress on medical informatics (MedInfo 2004)* (ed Fieschi M, Coiera E and Li YCJ), San Francisco, CA, 7–14 September 2004, pp. 716–720. Amsterdam: IOS Press.
12. Cohen G, Hilario M, Sax H, et al. Learning from imbalanced data in surveillance of nosocomial infection. *Artif Intell Med* 2006; 37(1): 7–18.
13. Iavindrasana J, Cohen G, Depeursinge A, et al. Minimal set of attributes required to report hospital-acquired infection cases. In: *Proceedings of the workshop on intelligent data analysis in biomedicine and pharmacology (IDAMAP 2008)* (ed Holmes J and Tucker A), Washington, DC, 7 November 2008, pp. 23–28.
14. Iavindrasana J, Cohen G, Depeursinge A, et al. Towards an automated nosocomial infection case reporting-framework to build a computer-aided detection of nosocomial infection. In: *Proceedings of the second international conference on health informatics (HEALTHINF 2009)* (ed Azevedo L and Londral AR), Porto, 14–17 January 2009, pp. 317–322. Porto: INSTICC Press.
15. Kelly KN and Monson JRT. Hospital-acquired infections. *Surgery* 2012; 30(12): 640–644.
16. Dalal MK and Zaveri MA. Automatic text classification: a technical review. *Int J Comput Appl* 2011; 28(2): 37–40.
17. Colas F and Brazdil P. Comparison of SVM and some older classification algorithms in text classification tasks. In: Bramer M (ed.) *IFIP international federation for information processing, vol. 217: artificial intelligence in theory and practice*. Boston, MA: Springer, 2006, pp. 169–178.
18. Hastie T, Tibshirani R, Friedman J, et al. The elements of statistical learning: data mining, inference and prediction. *Math Intell* 2005; 27(2): 83–85.
19. Noble WS. What is a support vector machine? *Nat Biotechnol* 2006; 24(12): 1565–1567.
20. Hsu CW, Chang CC and Lin CJ. A practical guide to support vector classification, https://www.cs.sfu.ca/people/Faculty/teaching/726/spring11/svmguide.pdf (2000, accessed 17 January 2014).
21. Yang Y and Pedersen JO. A comparative study on feature selection in text categorization. In: *Proceedings of the fourteenth international conference on machine learning (ICML'97)* (ed Fisher DH), Nashville, TN, 8–12 July 1997, pp. 412–420. San Francisco, CA: Morgan Kaufmann Publishers Inc.
22. Manning CD, Raghavan P and Schütze H. *Introduction to information retrieval* (online edition). Cambridge: Cambridge University Press, 2009, p. 27.
23. Van Rijsbergen CJ. *Information retrieval*. 2nd ed. London: Butterworth, 1979, p. 208.
24. Dragut E, Fang F, Sistla P, et al. Stop word and related problems in web interface integration. In: *Proceedings of the VLDB endowment* (ed Jagadish HV), Lyon, 24–28 August 2009, vol. 2, pp. 349–360. New York: ACM.

25. Hassel M. JavaSDM: a Java package for working with Random Indexing and Granska, http://www.nada.kth.se/~xmartin/java/ (2006, accessed 17 January 2014).
26. Doraisamy S, Golzari S, Norowi NM, et al. A Study on feature selection and classification techniques for automatic genre classification of traditional malay music. In: *Proceedings of the 9th international conference of music information retrieval (ISMIR 2008)* (ed Bello JP, Chew E and Turnbull D), Philadelphia, PA, 14–18 September 2008, pp. 331–336. Drexel University.
27. Japkowicz N and Shah M. *Evaluating learning algorithms: a classification perspective*. 1st ed. Cambridge: Cambridge University Press, 2011, p. 231.
28. Tegnell A and Carlson J. Att förebygga vårdrelaterade infektioner. Ett kunskapsunderlag, https://www.folkhalsomyndigheten.se/pagefiles/20412/att-forebygga-vardrelaterade-infektioner-ett-kunskapsunderlag-2006-123-12.pdf (2006, accessed 21 March 2015).
29. Bouzbid S, Gicquel Q, Gerbier S, et al. Automated detection of nosocomial infections: evaluation of different strategies in an intensive care unit 2000–2006. *J Hosp Infect* 2011; 79(1): 38–43.
30. Hall MA. *Correlation-based feature selection for machine learning*. PhD Thesis, The University of Waikato, Hamilton, New Zealand, 1999.

Article

Health Informatics Journal

Health Informatics Journal
2018, Vol. 24(1) 43–53
© The Author(s) 2016
Reprints and permissions:
sagepub.co.uk/journalsPermissions.nav
DOI: 10.1177/1460458216654288
journals.sagepub.com/home/jhi
SAGE

Infrastructures for healthcare: From synergy to reverse synergy

Tue Odd Langhoff, Mikkel Hvid Amstrup and Peter Mørck
IT University of Copenhagen, Denmark

Pernille Bjørn
University of Copenhagen, Denmark

Abstract

The *Danish General Practitioners Database* has over more than a decade developed into a large-scale successful information infrastructure supporting medical research in Denmark. Danish general practitioners produce the data, by coding all patient consultations according to a certain set of classifications, on the entire Danish population. However, in the Autumn of 2014, the system was temporarily shut down due to a lawsuit filed by two general practitioners. In this article, we ask why and identify a political struggle concerning authority, control, and autonomy related to a transformation of the fundamental ontology of the information infrastructure. We explore how the transformed ontology created cracks in the inertia of the information infrastructure damaging the long-term sustainability. We propose the concept of reverse synergy as the awareness of negative impacts occurring when uncritically adding new actors or purposes to a system without due consideration to the nature of the infrastructure. We argue that while long-term information infrastructures are dynamic by nature and constantly impacted by actors joining or leaving the project, each activity of adding new actors must take reverse synergy into account, if not to risk breaking down the fragile nature of otherwise successful information infrastructures supporting research on healthcare.

Keywords

data security and confidentiality, databases and data mining, electronic health records, information and knowledge management, information infrastructures, primary care, quality control, sustainability, synergy

Introduction

Healthcare in Denmark is a large, complex, public-funded, distributed organization consisting of 54 hospitals, 3510 general practitioners, and 901 specialized medical doctors servicing 5.6 million Danish citizens.[1] Creating integrated care across these healthcare institutions is a huge challenge,[2]

Corresponding author:
Peter Mørck, IT University of Copenhagen, Rued Langgaards Vej 7, Copenhagen S 2300, Denmark.
Email: pemo@itu.dk

not only in Denmark but also in all the Nordic countries. Research reveals the distinct fragmentation across technical, organizational, and professional boundaries within and across healthcare institutions.[3] In particular, we find documentation of the difficulties in establishing and maintaining information infrastructures supporting healthcare,[4,5] such as the challenges of introducing standardizations while taking into account contextual contingencies through reconfiguration,[6,7] as a way to reduce resistance toward changes.[8] Often, these studies are concerned with the difficulties in making connections and relations across the various parts of the healthcare organization, which are particularly complex due to the distributed and interdisciplinary organization. The *Danish General Practitioners Database* is an example of an information infrastructure which succeeded in creating connections and relations across multiple otherwise fragmented entities of the Danish primary healthcare sector. The *Danish General Practitioners Database* was created more than 10 years ago as a tool to collect information about all patients in the Danish society visiting their general practitioner and then use the information to learn and develop new best practices for care. The number of individual information technology (IT) systems in general practice is currently 11;[9] however, every practice can choose their own IT system among these. The *Sentinel* data capture module was thus created as an add-on system collecting data from all the individual systems connecting the many geographically dispersed general practices into the *Danish General Practitioner Database*. Since 2001, the general practitioners have been able to install and integrate the special data capture module, *Sentinel*, into their own medical systems supporting production, collection, and comparison of data across all the general practitioners. Over time, as participation increased, new changes were made to the database, for example, the diagnosis standard ICPC-2 was implemented.[10] The database managed to do, what other systems have failed, namely to be implemented and used in many independent clinics around the country collecting data about the whole population of Denmark. The database is extremely valuable for medical research, particularly in the area of the four chronic diseases: diabetes, depression, congestive heart failure, and chronic obstructive pulmonary disease. However, the successful story of the *Danish General Practitioners Database* ended in the Autumn of 2014, where it was temporarily shut down and today more than 1 year later *no* new data are being entered into the system.[9] This situation was caused by a lawsuit that challenged the legitimacy of the fundamental structure of the database and was filed by two general practitioners. It had come to their attention that since 2007 general practitioners classified patient data had been illegally collected and shared among *other* parties in the Danish healthcare sector, without the consent of the patients or the general practitioners, violating the Danish Penal Code, Privacy-Data Act and Health Act.[11] While this lawsuit is still unsettled, the database is not in use; and we ask why? What is the problem with the database? Why have it been decided necessary to shut down an otherwise successful and impressive information infrastructure producing valuable data to be used for research on health?

In this article, we examine the empirical case of the shutdown of the *Danish General Practitioners Database* as an example of particular socio-technical challenges of politically contested connections embedded within an information infrastructure for healthcare. How come this successful infrastructure connecting entities, research, and domains prove so problematic? In particular, we explore the transformation of the database from an innovation in improving medical practice toward a politically contested infrastructure. Building upon current research on information infrastructures, we argue that what makes the database politically contested lies in the ontological transformation when new actors were included as users with agendas clashing with existing participants. Interestingly, we found that the core reason behind the lawsuit was partly grounded in concerns about the privacy of the patients, but also partly as a reaction to the change of the basic ontology of the system. The more successful the database became, the more new entities and partners saw potential in joining the "success" of the database, introducing their own agendas. This transformation broke the core inertia of the infrastructure, which led to the shutdown. Clearly, the current stage of the information infrastructure is not sustainable; instead, we have a situation of *reverse synergy.* Our case demonstrates that in certain cases under particular circumstances continuous inclusion of new partners is not always a good strategy, since even though the perceived benefits might seem clear, it can risk jeopardizing the fundamental ontology of the information infrastructure, which can lead to a breakdown.

Method and empirical data

As phrased by Leigh Star, studying infrastructures is about studying the "boring things."[12] It is about unpacking and digging underneath the technical system design to figure out and restore the narratives, which are inherently embedded within the information infrastructure. When we first began exploring the *Danish General Practitioners Database*, we wanted to figure out how the technical system went from being an important innovation improving medical research and practice toward a politically contested infrastructure with a lawsuit. We took a multi-sited ethnography approach[13–15] and followed the essential links, connections, and associations involved in the case. Our focus is the Danish primary healthcare sector, which includes 3510 general practitioners, who are private enterprises, however, publically funded through universal healthcare for the Danish population. In this way, we were able to explore the information infrastructure as a "multiplicity"[16] and thus moved beyond "locality" of sites as spatially and temporally bounded.[13] Clearly, the information infrastructure was not bound by spatial or temporal constraints, but instead comprised a structure, which was continuously transformed over time by the multiple actors. We wanted to explore how a successful information infrastructure's progression over a whole decade "died" because it was "too successful." By conducting multi-sited ethnography, we examined multiple information sources derived from different spatial and temporal settings, which allowed us to investigate the transformation of actors and technical specifications that made up the basic nature of the *Danish General Practitioners Database*.

Our journey made us follow the information infrastructure into the general practitioner's office, outside the office and into the medical research scene, and from the scene of research into the political scene of politicians and unions. In total, we conducted eight interviews with key actors. Prior to the interviews, each interviewee was introduced to the research question and gave their consent to participate non-anonymized. So in this case all interviewees knew the exact scope and content of what we were trying to do. Interviews included three Danish general practitioners—here two interviews focused on the practices, while one interview was with one of the two general practitioners who had placed the lawsuit. From the general practice, we also wanted to explore the technical aspects of the infrastructure and interviewed the director of the Danish General Practitioners Quality Unit (DAK-E), who is located in Odense, Denmark, and responsible for managing the database. The interview with the director made it clear to us that if we were to understand the complexity of the case, we had to explore the political scene. Therefore, we interviewed the chairman of the Danish Regions Salary and Fare Committee. His role was closely connected to the negotiations of the settlement agreements between the Danish regions and the General Practitioners Organization. In this interview, the chairman explained and expressed the case from the Danish regions perspective. Continuing to unpack the political scene, we interviewed the at the time (now former) director and chief economist for the Danish regions, who had also been deeply involved in the case. In Denmark, the Danish regions are the main governmental entity in charge of Danish healthcare, but are also responsible for handling tasks within social services and regional development. In this interview, it became clear to us how the role of the *Danish General Practitioners Database* was to include servicing the political agenda of controlling the work of the otherwise "independent" general practitioners work practices. It is important to notice that in Denmark, while healthcare is universal and public, the general practitioners are organized as private entities, paid for their treatments by the government. Moving out of the general practice and the political scene, we then went to explore the research scene involved. We interviewed the vice director of the State Serum Institute. The State Serum institute is a public organization mainly concerned with coordinating, among others, IT support in healthcare and has responsibility for the operation and development of healthcare IT systems within the Danish Ministry of Health. We also wanted to include the private research interests of the *Danish General Practitioners Database* and we therefore interviewed a chief consultant at the Danish Association of the Pharmaceutical Industry. Besides all the interviews, we also collected multiple documents and reports, as well as press articles and blog entries related to the case, this included documents on Danish law and legislation in correlation with memos and records from previous and current settlement agreements.[i]

Infrastructuring ontologies, sustainability, and synergy

Exploring the transformation of the information infrastructure of the *Danish General Practitioners Database*, we identified three important aspects, which assisted us in unpacking the complexity of our case, namely: ontology, sustainability, and synergy.

Ontologies are important aspects of infrastructures. When exploring information infrastructures, it is important to acknowledge that information, and therefore infrastructures, do not exist as something outside social worlds, but rather as something that has a specific purpose and context based upon certain ontological conceptualizations.[17] Ontologies refer to the models or classification systems, which serve as the foundation for healthcare information systems. The ontology of an information system includes the interrelationships between entities structured within a hierarchy including subdivision of similarities and differences across data. The hierarchy is based upon the value set, which make pertinent the important relationships across entities. The practical application of an ontology is the taxonomy by which an information system is designed. In this article, we refer to two different types of ontologies: the research agenda ontology and the governmental agenda ontology. It is important to notice that while the technical design of the healthcare information system (the taxonomy) remains the same, what is different is the ontological comprehension of the system, which divert. We can say with the words of Mol[16] that the ontology of the *Danish General Practitioners Database* is not given in the order of things, but instead "ontolog*ies* are brought into being, sustained, or allowed to wither away in common, day-to-day, sociomaterial practices." We find that the research agenda ontology is based on the value scheme of quality in research on medical practices, while the governmental agenda ontology is based on the value scheme for control and expanding the quality assurance to a wider national perspective. As such, what we conceptualize as an ontology is not simply discourses or social worlds but fundamental ideas of *what counts.* However, ontologies are not stable entities, instead they are dynamically changing in-use, and practices which are part of defining the ontologies come from the practices in-use by which the infrastructure becomes *infrastructured.*[18] *Infrastructuring* is the practices by which the infrastructure is enacted and how this enactment shapes the nature of the infrastructure. Ontologies stipulate the classification of what can be included and what is excluded and thus makes the inertia of the infrastructure. Hence, ontology building is an important activity when designing IT systems for data collection and is basically the act of systematizing knowledge from a certain domain defined by experts and allowing for reuse of the information by different fields.[19] Interestingly, by allowing for reuse by other actors, the ability for others to take part in actively reshaping the data is also introduced. Decisions on what to include or exclude is not based on clear-cut categorization; instead, participants continuously negotiate these decisions. In healthcare, there are several examples of how important decisions about categorizations are taken without including the healthcare practitioners who are most affected by it, for example, categorization embedded within electronic triage systems.[20,21] When we examine the transformation of the *Danish General Practitioners Database*, we must explore the ways in which *infrastructuring activities* in practice *transform the ontology* of the categorization, keeping in mind the multiple layers and complexities of the infrastructure.[8,22] Exploring the transformation of the ontology in the database, we must pay attention to the multiple practices of infrastructuring which become embedded within the categorization and how these manage or fail to coexist across different social worlds and agendas.

The *sustainability* of an information infrastructure draws on many resources both technical and social to evolve and endure—craving for perpetual actions to be taken.[23] Furthermore, the sustainability of an information infrastructure is a continual process of maintaining infrastructural relationships—among people, organizations, and technologies—while withstanding temporal change.[24] In time an information infrastructure may increase in value as it connects to other systems and infrastructures,[24] subsequently creating interdependencies in-between local and global users and technologies.[25] To survive the wear and tear of time, an information infrastructure must evolve and adapt to environmental change and new demands to ensure its use and relevance, locally and globally.[24] Thus, in sustaining an infrastructure the distinction between local and global becomes morphed into one, and the interdependency in-between the two becomes blurred and difficult to separate.[25] Therefore, when we explore the ways in which the *Danish General Practitioners Database* become sustainable over time, we must examine the relation between the sustainability and the morphed dependencies, which emerge when new actors are introduced.

The notion of *synergy* is laden with the idea of positive alignment between actors.[26] Achieving synergy between multiple actors is important for designers and users of infrastructures in healthcare since synergy in systems connects dispersed actors increasing possibilities for collaboration. We will explore the ways in which *alignment work* and *leveraging* are produced in our case, paying attention to costs and benefits of such synergies for the information infrastructure. Furthermore, the process of *alignment work* and *leveraging* may be seen as a complex path of transfer from across diverse entities and ontologies. This process of transfer may affect technologies but also social, cultural, and organizational practices.[27] Additionally, transfer has a tendency to spawn conflicts, new challenges, and constituencies among its stakeholders.[27] This tension may be viewed as a consequence of an infrastructure bumping up against competing organizational, political, or financial objectives, or incompatible technologies.[27] Hence, the sustainability of an information infrastructure cannot be conceived in a vacuum decoupled from surrounding social, cultural, and historical events.[23] Thus, we explore the ways in which past and future plans of actions create conflicts, at present, in-between actors threatening the sustainability of *the Danish General Practitioners Database.*

Ontological misalignment, unsustainable infrastructure, and misaligned synergies

Ontological misalignment

Each time new actors joined the *Danish General Practitioners Database* over the years, the purpose and context was transformed. New actors took part in the infrastructuring practices and brought in their agendas. Initially, the enactment of the *Danish General Practitioners Database* concerned general practitioners providing health data for general practitioners with the sole agenda of improving in the professional context of general practice. For years, the main purpose of the infrastructure was to provide the general practitioners a complete overview of their patients and enable them to continuously receive quality assurance reports. As such, making it possible for general practitioners to evaluate and improve not only their practice as whole but also the treatment quality of each patient. As explained by the director of the database,

> What we do at Dak-E is quality development, with support for the patients but also for the general practitioners […] The most important quality development take place in general practice […] Our goal is to support this quality development through data. (Director of Dak-E, 15 April 2015)

However, this initial purpose changed as the health data were deemed valuable for broader research purposes than simply improving the general practice. The value was especially connected to the fact that the *Danish General Practitioners Database* became a place where general practitioners collected and stored health data about the complete Danish population. The broader medical field hence recognized a huge potential for doing new research, for example, in developing new medicine or knowing about the effect of special treatments. While the general practitioners, being medical professionals, could clearly see the value of opening up the data for research, they also raised concerns about patient confidentiality as well as possible misuse; who would gain access to the data? What would the data be used for? After some debate surrounding these concerns, the final decision was to open up access to the data for broader research; however, the pharmaceutical industry was kept out entirely. Access to the health data was only given to less than a hundred smaller research projects on an aggregated level and only after the research project had gone through several validation steps first.[28] After the addition of new actors, in the form of researchers, to the infrastructure, the infrastructuring practices of the database were transformed; however, the *basic ontological structure remained the same*. Thus, the general practitioners experienced no problematic changes to their daily practice of collecting data. After granting the research projects' access to the data, the success and further expansion of the information infrastructure continued for several years, until a new actor wanted to be included. The new actor was the governmental entity of the Danish regions, which saw an opportunity to join the successful infrastructure for two reasons. First, they wanted to create a more integrated primary and secondary healthcare sector allowing an overall better treatment experience for the patients across the distributed healthcare organization in Denmark. Second, they wanted to gain more control over a somewhat autonomous primary care sector (general practitioners in Denmark) and obtain a tool, which could assist them to more closely monitor and control that work was done in accordance with protocol. While no one in Denmark would contest the first agenda, the second agenda turned out to be highly contested. As the former chief economist for the Danish regions expresses,

> In their own perception, they [general practitioners] are the lords of their own mansion. However, it's a misconception, when they believe no one controls them […] this is not how you run a healthcare sector, and especially not a public funded healthcare system as the Danish one. We [Danish citizens and government] need to know what we are getting for our money […]. (Former director and chief economist for the Danish Regions, 8 April 2015)

In the Danish regions attempt to gain access to the data, and in this process also actively changing the design and use of the *Danish General Practitioners Database*, to accommodate their agendas, they initiated negotiations with the General Practitioners Organization. In these negotiations, the Danish regions pushed to create a settlement agreement which instructed that it would be mandatory for general practitioners to provide information accordingly to the IT system, including the additional fields for data entry made by the regions to support their agenda of transparent governance. However, these negotiations broke down and the Danish regions in collaboration with the Danish state decided to make data capture mandatory by law instead and hence the general practitioners could not directly contest or negotiate the change. So where it was previously a choice for the general practitioners to record the data on their patients, it was now written into the law regarding the procedures for general practice.[29] With this legal change to the information infrastructure, the ontology of the information infrastructure was altered in crucial ways, since the purpose, use context, and infrastructuring practices were transformed dramatically. By introducing legal entities into the infrastructure, the ontology was no longer for research, since by inviting in new actors they become part of reshaping the nature of the data and thus the ontology.[19] Before the legal change, the ontology included classification *only* for healthcare research purposes supporting improvement of medical practices in the general practitioners' domain and other aggregated research projects. After the legal change, the ontological structure of the database was now stretched to include governing and financial agendas. Where the infrastructure used to be for bottom-up research data collection for peers, it now emerged as a top-down governmental control mechanism with direct financial impact.

Unsustainable infrastructure

Numerous social, technical, and financial resources have been allocated to develop and sustain the *Danish General Practitioners Database's* information infrastructure.[10] As such, long-term sustainability of the infrastructure was key already at the initiation of the system. As discussed earlier, the information infrastructure has continuously transformed over time and survived expansions and adaptations related to research. Clearly, it was not shut down due to technical obstacles or technical integration across patchwork systems, which have been reported as reasons for lack of sustainability.[30] When the ontology changed due to the introduction of a governance agenda to the system, the *research* data *transformed* into *governance* data, which shook the inertia of the installed base.[18] The economic and financial restrictions now introduced to the system, impacted how the data were recorded, and thus also the *basic nature* of the data and the infrastructure. As a general practitioner explains,

> I am sure there are errors [in the data]. It needs to go fast [when recording health data], as I cannot sit and philosophize over patients with for example diabetes, and then choose between 20 categories with or without related eye symptoms, it is not realistic […]. (General Practitioner, #1, 25 February 2015)

That healthcare data, through the introduction of new electronic systems introduces multiple data use purposes, is not a new phenomenon[31] and prior research have documented problems of how secondary use of data impact the primary purpose of data collection.[20] Interestingly, the synergy between research data and governance data was imagined as unproblematic by the Danish regions. They assumed that by joining the in-use information infrastructure for research data, they were able to get a tool for administrative governance of the distributed healthcare organization of the general practitioners. Changing the infrastructure by changing the legal foundation was seen as straightforward. The Danish regions used legal measures to join the infrastructure, and in turn the general practitioners used legal measures to fight this fundamental change of the systems' ontology, subsequently shutting down the data capture activities. The growing tensions encountered by the general practitioners created skepticism toward the system. As described by a general practitioner,

> It's not like I don't want order in my work, it should be done well and thoroughly, but the point is that it to some extent must be realistic and we are now in a situation where it in my opinion have lost its sense of reality which makes it [recording health data] untrustworthy and you lose faith in it, unfortunately, and then it pales that it could actually be good for quality assurance. (General Practitioner, #1, 25 February 2015)

Adapting new systems always include extra activities such as learning new coding schemes, standards, and technology-in-use practices.[32] However, frictions toward the technical aspects of the system grew proportionally with the discussions of reusing data in different social worlds outside healthcare research. This was not a clash of interests derived from transferring technology or protocol from one domain to another; instead, it sparked from the general practitioners' position to protect the integrity of the health data, the privacy of their patients, and the autonomy of their general practice from other interest groups and agendas. Allegedly, a growing number of general practitioners purposely began to miscode patient data cloaking it to protect their patients and practices from possible misuse. Interestingly, prior research have also found how healthcare practitioners purposefully use other types of "coding" in free text forms, when the IT systems risk jeopardizing the work flow of patients—for example, by writing "lost 20 kg" rather than suspicion of cancer.[33] The issue expressed by a general practitioner,

> I am very reluctant when I fill in something and for example give the diagnose depression then I might write stress, overloaded, sad, or have used 27 handkerchiefs, instead. (General Practitioner, #2, 26 February 2015)

The sustainability of the *Danish General Practitioners Database* information infrastructure became threatened by the growing uncertainty about context and the purpose of the database. Even if the governance agendas were removed from the system, the general practitioners would still be conscious in how they record data, since *new* top-down initiatives might be introduced in the future. When political agendas can force the transformation of the ontology of the infrastructure through law and regulations in this case, then it can be repeated again in other situations. This uncertainty jeopardizes the sustainability of the infrastructure because of potential future initiatives and motives, which might emerge and be forced upon the general practitioners.

Misaligned synergies

Synergy was not achieved. Instead, the general practitioners began to combat the new mandatory ontological understanding of the data for governance. In turn, this battle caused stagnation for the information infrastructure. Stagnation happens, when people resist required adaption, due to technological, legislative, or social reconfigurations, by referring to the ways in which the system privileges one professional group over another.[3] In our case, the loss of trust in the infrastructure caused by the legal changes made general practitioners create workarounds in their use of the system. To safeguard the privacy of information regarding their patients ensuring confidentiality, they omitted critical information or wrote it in non-searchable free text rather than selecting pre-defined classifications and diagnoses. As a general practitioner describe it,

> If a person has an alcohol problem we can always just give him his medicine under the table. (General Practitioner, #3, 9 March 2015)

Workarounds were not developed due to time constraints or changed procedures of the IT system, but rather as a deliberate act. By documenting information in this way, or by simply omitting it completely, the general practitioners—dissatisfied with the government's access to what they perceived as their confidential data—successfully worked around the governance agenda. When the general practitioners were fighting against the information infrastructure through workarounds; the data which theoretically could serve multiple purposes, ended up with ironically not being able to serve *either*. We thus argue that creating synergy between the general practitioners and the Danish regions, using the *Danish General Practitioners Database* as a tool for sharing and collecting data, is not a viable approach and conceivably the main reason for the dispute. It was not possible for the information infrastructure to bridge across conceptually diverse social worlds, which fundamentally build upon very opposite ontological concerns (research data or governance data). Forcing synergy thus risks losing important tacit and contextual data for both parties. In this way, insisting on synergy the very nature and purpose of what both partners try to achieve might be lost. We have a lockdown situation of reverse synergy within the information infrastructure. *Reverse synergy is the effect when the effort needed to align diverse actors creates enough cracks in the inertia in a given information infrastructure, erodes enough social capital, or in other ways require an amount of alignment work, articulation work, or other types of coordination that outweighs the additional value gained from including the actors in the first place.* To mitigate the negative effects of the reverse synergy, which have emerged in the dispute of the *Danish General Practitioners Database*, it might require less integration rather than more alignment. Our empirical data demonstrate that even though the collaboration across different social worlds and fundamentally different ontologies might theoretically create *additional value for all*; in some cases, it will create *no value at all*.

Reverse synergy

While we agree with Berg,[34] that what counts as success in healthcare information infrastructures is tricky to define; we would argue that the *Danish General Practitioners Database* was a successful information infrastructure, which then ended up as a total failure due to reverse synergy. Our concept of reverse synergy is challenging the dominant idea in contemporary information infrastructure literature concerning healthcare, where studies continuously argue how adding new actors adds new potential value, and where failures are often seen as misalignment between actors, which could have been mitigated.[4,5] We argue that in our case the situation was not about lack of alignment, but rather the underlying concern held by core actors related to the uncertainty for how fundamentally incompatible agendas were introduced. Our case demonstrates that connecting new actors uncritically to an information infrastructure can have serious consequences for the existence and further development of it. Clearly, information infrastructures transform over time, and we do not argue for isolation and closure. Rather, we bring to the surface how potential conflicts about the ontological structure of an infrastructure introduced through political measures; encourage key participants to resist. In our case, resistance came in terms of a lawsuit and local workarounds fighting what was seen as political strong agendas forced top-down, but in other cases resistance might take different forms, which is why future research should investigate similar cases in other parts of the world such as the United Kingdom[35] and Scotland.[36] In turn, by comparing and analyzing multiple cases, we might then be able to identify strategies of how to avoid reverse synergy. Our purpose is not to suggest strategies for reducing risks of resistance. Rather, our case demonstrates costs of reverse synergy being corrosive to the ontological inertia initially built into the infrastructure. Our article unpacks the oft-missed cost of misaligned actors and conveys the fundamental idea that adding new actors, agendas, or purposes risk having serious consequences for existing successful information infrastructures.

Acknowledgements

Earlier versions of this paper were presented at the 5th International workshop on ‘Infrastructures for Health Care: Patient-Centered Care and Patient Generated Data’ in Trento June 2015; and at the ‘Knowledge Infrastructure’ workshop held in Seattle in December 2015. We are grateful to all participants who have made comments and especially to David Ribes, Ola Hanseth, and Stephanie Steinhardt for their comments and discussion. We would also like to thank the anonymous reviewers for insightful comments and suggestions as well as the people who lent us their time to be interviewed for this paper. Finally, we want to extend a special thanks to David Randall for discussing the initial scope of the project over a cup of coffee and to Charlotte Lee for her constructive comments and contribution in critically discussing the argument in earlier versions of this paper.

Declaration of Conflicting Interests

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding

The author(s) received no financial support for the research, authorship, and/or publication of this article.

Notes

i. Documents include the following:

1. OECD (2013) OECD reviews of health care quality: Denmark. Available at: http://www.oecd.org/els/health-systems/ReviewofHealthCareQualityDENMARK_ExecutiveSummary.pdf (accessed 13 April 2016).
2. Kristensen TG (n.d.) DAMD-opfinder: Striden er vel konsekvensen af vores succes. Available at: http://www.altinget.dk/artikel/damd-databasens-opfinder-det-er-vel-konsekvensen-af-vores-succes
3. Lægeforeningen (n.d.) Yngre læger og PLO-forhandlingerne. Available at: http://www.laeger.dk/portal/page/portal/LAEGERDK/Laegerdk/Y_L/Overenskomst/ALMEN_PRAKSIS/
4. Practicus (2013) Data til kvalitetsudviklings- og forskningsprojekter fra Dansk AlmenMedicinsk Database (DAMD). Available at: http://www.practicus.dk/flx/artikler/?m=showArticle&aid=243
5. Rasmussen LL (2006). Bekendtgørelse om indberetning af oplysninger til kliniske kvalitetsdatabaser m.v, 21 December. Available at: https://www.retsinformation.dk/Forms/R0710.aspx?id=11046

References

1. Herning L. Fakta om sundhedsvæsenet - sundhedsvæsenet i tal. *Regioner*, 3 June, http://regioner.dk/Aktuelt/Temaer/Fakta+om+regionernes+effektivitet+og+%C3%B8konomi/Kopi+af+Fakta+om+sundhedsv%C3%A6senet.aspx (2015, accessed 10 November 2015).
2. Andersen T, Bjørn P, Kensing F, et al. Designing for collaborative interpretation in telemonitoring: re-introducing patients as diagnostic agents. *Int J Med Inform* 2011; 80(8): e112–e126.
3. Ellingsen G and Munkvold G. Infrastructural arrangements for integrated care: implementing an electronic nursing plan in a psychogeriatric ward. *Int J Integr Care* 2007; 7(2): e13.
4. Lecluijze I, Penders B, Feron F, et al. Infrastructural work in child welfare: incommensurable politics in the Dutch child Index. *Scand J Inform Syst* 2014; 26(2): 31–52.
5. Modol J and Chekanov A. Architectural constraints on the bootstrapping of a personal health record. *Scand J Inform Syst* 2014; 26(2): 53–78.
6. Meum T, Monteiro E and Ellingsen G. The pendulum of standardization. In: *Proceedings of the 12th European conference on computer supported cooperative work*, Aarhus, 24–28 September 2011.
7. Bjørn P, Burgoyne S, Crompton V, et al. Boundary factors and contextual contingencies: configuring electronic templates for healthcare professionals. *Eur J Inform Syst* 2009; 18(5): 428–441.
8. Hanseth O, Monteiro E and Hatling M. Developing information infrastructure: the tension between standardization and flexibility. *Sci Technol Hum Val* 1996; 21(4): 407–426.
9. DAK-E. Datafangst. *Dak-e*, http://www.dak-e.dk/praksis/datafangst.php (2016, accessed 25 April 2016).
10. Schroll H. Historie. *Dak-e*, http://www.dak-e.dk/dake/historie.php (2016, accessed 25 April 2016).
11. Rasmussen E. Læger politianmelder Region Syddanmark for ulovlig indsamling af patientdata. *Politikken*, 12 November, http://politiken.dk/forbrugogliv/forbrug/forbrugersikkerhed/ECE2451830/laeger-politian-melder-region-syddanmark-for-ulovlig-indsamling-af-patientdata/ (2014, 11 March 2015).
12. Star S. The ethnography of infrastructure. *Am Behav Sci* 1999; 43(3): 377–391.

13. Marcus G. Ethnography in/of the world system: the emergence of multi-sited ethnography. *Annu Rev Anthropol* 1995; 24: 95–117.
14. Blomberg J and Karasti H. Reflections on 25 years of ethnography in CSCW. *Comp Support Coop W* 2013; 22(4–6): 373–423.
15. Bjørn P and Boulus-Rødje N. The multiple intersecting sites of design in CSCW research. *Comp Support Coop W* 2015; 24(4): 319–351.
16. Mol A. *The body multiple: ontology in medical practice*. Durham, NC: Duke University Press, 2002.
17. Ribes D and Polk J. Historical ontology and infrastructure. In: *Proceedings of the 2012 iConference*, Toronto, CA, 7 February 2012, pp. 254–262. New York: ACM.
18. Bowker G and Star S. *Sorting things out: classification and its consequences*. Cambridge, MA: The MIT Press, 2000, pp. 1–51.
19. Bowker G, Baker C, Millerand F, et al. Toward information infrastructure studies: ways of knowing in a networked environment. In: Hunsinger J (ed.) *International handbook of internet research*. Rotterdam: Springer, 2010, pp. 97–117.
20. Bjørn P and Balka E. Health care categories have politics too: unpacking the managerial agendas of electronic triage systems. In: Bannon L, Wagner I, Gutwin C, et al. (eds) *ECSCW 2007*. London: Springer, pp. 371–390.
21. Ash J, Berg M and Coiera E. Some unintended consequences of information technology in health care: the nature of patient care information system-related errors. *J Am Med Inform Assn* 2004; 11(2): 104–112.
22. Matthiesen S and Bjørn P. Why replacing legacy systems is so hard in global software development: an information infrastructure perspective. In: *Proceedings of the 18th ACM conference on computer supported cooperative work & social computing*, Vancouver, BC, Canada, 28 February 2015, pp. 876–880. New York: ACM.
23. Braa J, Monteiro E and Sahay S. Networks of action: sustainable health information systems across developing countries. *Mis Quart* 2004; 28: 337–362.
24. Bietz M, Ferro T and Lee C. Sustaining the development of cyberinfrastructure: an organization adapting to change. In: *Proceedings of the ACM 2012 conference on computer supported cooperative work*, Seattle, WA, 11–15 February 2012, pp. 901–910. New York: ACM.
25. Ellingsen G, Monteiro E and Røed K. Integration as interdependent workaround. *Int J Med Inform* 2013; 82(5): e161–e169.
26. Bietz M and Lee C. Adapting cyberinfrastructure to new science: tensions and strategies. In: *Proceedings of the 2012 iConference*, Toronto, ON, Canada, 7–10 February 2012, pp. 183–190. New York: ACM.
27. Jackson SJ, Edwards PN, Bowker GC and Knobel CP. Understanding infrastructure: history, heuristics and cyberinfrastructure policy. *First Monday* 12(6), 4 June 2007, http://firstmonday.org/issue/view/240
28. DAK-E. Kvalitets- og forskningsudvalget. *Dak-e,* http://dak-e.dk/damd/kvalitetsogforskningsudvalg.php (2016, accessed 25 April 2016).
29. Hækkerup N. Sundhedsloven. *Retsinformation*, 26 December, https://www.retsinformation.dk/Forms/r0710.aspx?id=152710 (2013, accessed 10 March 2015).
30. Ellingsen G and Monteiro E. A patchwork planet integration and cooperation in hospitals. *Comp Support Coop W* 2003; 12(1): 71–95.
31. Balka E and Whitehouse S. Whose work practice? Situating an electronic triage system within a complex system. *St Heal T* 2006; 130: 59–74.
32. Boulus-Rødje N and Bjørn P. A cross-case analysis of technology-in-use practices: EPR-adaptation in Canada and Norway. *Int J Med Inform* 2010; 79(6): e97–e108.
33. Møller N and Bjørn P. Layers in sorting practices: sorting out patients with potential cancer. *Comp Support Coop W* 2011; 20(3): 123–153.
34. Berg M. Implementing information systems in health care organizations: myths and challenges. *Int J Med Inform* 2001; 64(2): 143–156.
35. Lusignan S and Weel C. The use of routinely collected computer data for research in primary care: opportunities and challenges. *Fam Pract* 2005; 23(2): 253–263.
36. Spire. Scottish Primary Care Information Resource. *Spire*, http://www.spire.scot.nhs.uk/ (2016, accessed 25 April 2016).

Health Informatics Journal

Article

Health Informatics Journal
2018, Vol. 24(1) 54–65
© The Author(s) 2016
Reprints and permissions:
sagepub.co.uk/journalsPermissions.nav
DOI: 10.1177/1460458216655188
journals.sagepub.com/home/jhi
SAGE

Logic Learning Machine and standard supervised methods for Hodgkin's lymphoma prognosis using gene expression data and clinical variables

Stefano Parodi
National Research Council of Italy, Italy

Chiara Manneschi
Italian Institute of Technology, Italy

Damiano Verda and Enrico Ferrari
Rulex Inc, USA

Marco Muselli
National Research Council of Italy, Italy

Abstract

This study evaluates the performance of a set of machine learning techniques in predicting the prognosis of Hodgkin's lymphoma using clinical factors and gene expression data. Analysed samples from 130 Hodgkin's lymphoma patients included a small set of clinical variables and more than 54,000 gene features. Machine learning classifiers included three black-box algorithms (*k*-nearest neighbour, Artificial Neural Network, and Support Vector Machine) and two methods based on intelligible rules (Decision Tree and the innovative Logic Learning Machine method). Support Vector Machine clearly outperformed any of the other methods. Among the two rule-based algorithms, Logic Learning Machine performed better and identified a set of simple intelligible rules based on a combination of clinical variables and gene expressions. Decision Tree identified a non-coding gene (*XIST*) involved in the early phases of X chromosome inactivation that was overexpressed in females and in non-relapsed patients. *XIST* expression might be responsible for the better prognosis of female Hodgkin's lymphoma patients.

Keywords

artificial neural network, cancer prognosis, Decision Tree, Hodgkin's lymphoma, Logic Learning Machine, Support Vector Machine

Corresponding author:
Stefano Parodi, Institute of Electronics, Computer and Telecommunication Engineering, National Research Council of Italy, Via De Marini, 6-16149 Genoa, Italy.
Email: parodistefano@icloud.com

Introduction

Hodgkin's lymphoma (HL) is a haematological malignancy accounting for about 10 per cent of all lymphoma cases in Western countries.[1,2] HL is composed of two distinct disease entities: classical HL, which accounts for about 95 per cent of the whole disease burden and is characterized by the presence of malignant multinucleated giant Reed–Sternberg cells, and nodular lymphocyte predominant HL, characterized by a neoplastic population of larger cells with folded lobulated nuclei.[3]

In the last decades, advances in radiation treatments and chemotherapy have greatly increased the survival rates of HL patients. Nonetheless, up to date, about 5–10 per cent of them are refractory to initial treatment and 10–30 per cent will relapse despite having achieved an initial complete remission.[4]

IPS (International Prognostic Score) is a prognostic index based on the combination of seven recognized prognostic factors for HL (namely, age ⩾45 years, stage IV, male sex, white blood count ⩾15,000 cells/mL, lymphocyte count <600 cells/mL, albumin <4.0 g/dL, haemoglobin <10.5 g/dL).[5] IPS was demonstrated to be predictive of the patient outcome in multivariable analysis. For instance, patients with five or more factors were found to have a 5-year progression-free survival of 42 per cent, while patients with non-negative prognostic factors had an 84 per cent probability of being free from progression at 5 years from diagnosis.[5] However, despite the quite good performance of IPS, the identification of new prognostic variables for HL patients is highly desirable to potentially increase patient survival and reduce treatment toxicity.[4] For this purpose, numerous studies have been carried out in the last few years and many new putative prognostic markers have been identified.[6,7] Among such studies, a large microarray experiment identified a set of 271 genes differently expressed between relapsed and non-relapsed patients.[8] Furthermore, the same study was able to associate a macrophage gene expression signature with primary treatment failure, even if this latter finding was questioned by further investigations.[9,10]

This study is aimed at evaluating the performance of a set of supervised machine learning techniques, including the recently proposed Logic Learning Machine (LLM) method, in predicting the prognosis of HL patients using clinical and gene expression data from the large data set by Steidl et al.[8]

Materials and methods

Database description

Data were downloaded from GDS4222.soft, a microarray database stored in the GEO repository[11] at http://www.ncbi.nlm.nih.gov/sites/GDSbrowser?acc=GDS4222. Data included information from 130 samples of classical HL and 54,675 gene expression features.[8]

Table 1 describes patient characteristics available in GDS4222.soft. Clinical and demographic variables included the following: relapse at any time after therapy (n = 38, 29.2%), gender (male: 56.2%; female: 43.8%), stage at diagnosis (Stage I: 12.3%; Stage II: 51.5%; Stage III: 22.3%; Stage IV: 13.8%) and IPS. This latter was aggregated into two categories, according to Steidl et al.[8]: high score, corresponding to IPS > 3 (24.6%), and low score, associated with IPS ⩽ 3 (75.4%). More details about patient selection, characteristics at diagnosis, assessment of disease status, primary line treatment, and methods for gene expression analysis, including data pre-processing and normalization, have been reported elsewhere.[8]

Supervised data mining methods

A set of supervised learning machine techniques was selected in order to predict HL patient prognosis. They included three common methods based on black-box algorithms (*k*-nearest neighbour

Table 1. Demographic and clinical characteristics of 130 patients with Hodgkin's lymphoma included in the analyses.

Patient characteristics	N	%
Gender		
Males	73	56.2
Females	57	43.8
Ann Arbor stage at diagnosis		
I	16	12.3
II	67	51.5
III	29	22.3
IV	18	13.8
International Prognostic Score		
Low (⩽3)	98	75.4
High (>3)	32	24.6
Follow up status after treatment		
Relapse	38	29.2
No Relapse	92	70.8

classifier, kNN; Artificial Neural Network, ANN; and Support Vector Machine, SVM) and two methods based on intelligible threshold-based rules (Decision Tree, DT, and the innovative LLM method). Standard classification based on the IPS score was also performed. The probability of relapse at any time after therapy was considered as the outcome, while all the available variables were used as input data. Accuracy measures included the total proportion of correctly classified samples (total accuracy) and the proportion of correct classifications among both relapsed (sensitivity) and non-relapsed patients (specificity).

In order to control the overfitting bias, accuracy estimates of each supervised analysis were obtained by cross-validation. Due to the rather small sample size of our data set, the leave-one-out procedure was adopted.[12] Finally, a comparison between the set of intelligible rules generated by LLM and DT was also performed.

All the analyses were carried out by using Rulex Analytics, a software suite developed and commercialized by Rulex Inc (http://www.rulex-inc.com).

kNN

Consider a training set S including n input–output pairs $(\boldsymbol{x}_j, y_j)$, with $j=1,\ldots, n$, where the output value y_j can be one of q possible classes, labelled by an integer A_i with $i=1,\ldots, q$. To classify any subject, described by an input vector $\boldsymbol{x}$, the nearest k samples (with respect to $\boldsymbol{x}$) in the training set S, according to a suitable distance measure, are considered. Then, the subject $\boldsymbol{x}$ is associated with the class A_i that characterizes the majority of the k-nearest samples.[13]

In the present investigation, the set of values $\{1, 3, 5\}$ was adopted for k and the standard Euclidean distance was employed, after having normalized the components of the input vector $\boldsymbol{x}$ to reduce the effect of biases possibly caused by unbalanced domain intervals in different input variables.

ANN

ANN is a connectionist model formed by the interconnection of simple units, called neurons, arranged in layers. The first layer receives the input vector $\boldsymbol{x}$, whereas the remaining layers receive their inputs from the previous one. Each neuron computes a weighted sum of the inputs and applies a proper activation function to obtain the output value that will be propagated to the following layer. The last layer produces the output class y to be assigned to $\boldsymbol{x}$. Weights for each neuron form the set of parameters for the ANN and are estimated by suitable optimization techniques.[13]

In this study, one intermediate layer was used, and the number of hidden neurons was allowed to vary from one to three. The nets were trained by means of the Levenberg–Marquardt version of the back propagation algorithm.[13]

SVM

SVM is a non-probabilistic binary linear classifier based on the identification of an optimal hyperplane of separation between two classes. Given a training set, the classifier selects a subset l of input vectors $\boldsymbol{x}_j$ in the training set S, called support vectors, and their corresponding outputs $y_j \in \{-1, 1\}$. The class y for any input vector $\boldsymbol{x}$ is given by

$$y = \operatorname{sgn}\left(\sum_{j=1}^{l} y_j \alpha_j K\left(\boldsymbol{x}_j, \boldsymbol{x}\right) + b\right)$$

where the coefficients α_j and the offset b are evaluated by the training algorithm.

$K(\cdot,\cdot)$ is a kernel function used to perform a non-linear classification by constructing an optimal hyperplane in a high dimensional projected space. Both a linear and a radial basis kernel function were tested on the GDS4222.soft data set. As it will be shown in the following section, in this case the linear kernel (which produces a linear classification) proves to be more robust with respect to overfitting. This is due to the fact that the classification problem is unbalanced (38 patients relapsed, while 92 did not), and moreover, the number of input attributes for classification far exceeds the number of training samples. The training algorithm was performed using the LIBSVM library, which is featured by the Rulex Analytics software.

DT

A DT is a graph where each node is associated with a condition based on an attribute of the input vector $\boldsymbol{x}$ (e.g. $x_i > 5$) and each leaf corresponds to an assignment for a specified output class. By navigating from a leaf to a root, a simple intelligible rule can be easily identified.[13] DT is generated by adopting a 'divide-and-conquer' approach that provides disjoint rules. At each iteration, a new node is added to the DT by choosing the condition that best subdivides the training set S according to a specific measure of goodness.

In the present investigation, the information gain I_G (also called 'the smallest maximum entropy') was employed as goodness indicator function. In more detail, given a set Q and a partition in q subsets $Q_1,\ldots, Q_q$, the information gain of Q with respect to the partition $\{Q,Q_j\}$ is defined by

$$I_G\left(Q,Q_j\right) = -\sum_{j=1}^{q} \frac{\left|Q_j\right|}{|Q|} log_2 \frac{\left|Q_j\right|}{|Q|}$$

where | . | indicates the number of elements in a set.

In our study, $q=1$ identifies the subset of non-relapsed patients and $q=2$ the subset of relapsed ones.

Finally, the pessimistic error pruning technique was adopted to reduce the complexity of the final DT and to increase its generalization ability. Briefly, let p be the error rate associated with a node s in a DT; all nodes and leafs below s are erased if the error p_{-s} associated with the node immediately below s exceeds the following quantity

$$+1.96\sqrt{\frac{p(1-p)}{n}}$$

where n represents the number of samples to be classified at the node r.[14]

LLM

LLM is an innovative method of supervised analysis based on an efficient implementation of the Switching Neural Network model,[15,16] which is associated with a classifier $g(\boldsymbol{x})$, described by a set of intelligible rules of the following type: **if** ‹*premise*› **then** ‹*consequence*›. The ‹*premise*› statement represents a logical product (AND) of conditions on the components of the input vector $\boldsymbol{x}$ and ‹*consequence*› provides a class assignment for the output y.

The general procedure employed to train an LLM passes through the following steps:

1. *Discretization*. Continuous and integer variables are properly discretized to reduce their variability, thus increasing the efficiency of the training algorithm and the accuracy of the resulting set of rules.
2. *Binarization*. Nominal and (discretized) ordered variables are coded into binary strings by adopting a suitable mapping that preserves ordering and distances.
3. *Logic synthesis*. Starting from the binarized version of the training set S, which can be viewed as a portion of a truth table, reconstruct the AND–OR expression of a consistent monotone Boolean function.
4. *Rule generation*. Transform every logical product of the AND–OR expression into an intelligible rule.

A valid and efficient way of performing Step 1 consists in adopting the attribute-driven incremental discretization (ADID),[17,18] which reduces the complexity of the input vector $\boldsymbol{x}$ while preserving the information included in the training set S concerning class discrimination. For each continuous or discrete input attribute, ADID is able to find a collection of separating points that lower its variability while maintaining its classification power. The core of ADID consists of an incremental algorithm that adds iteratively the cut-off scoring the highest value of a proper quality measure based on the capability of separating patterns of different classes. Smart updating procedures enable ADID to efficiently get an optimal discretization. Usually, ADID produces a minimal set of cut-offs for separating all the patterns belonging to different classes.[16]

Then, the (inverse) only-one coding[15] is adopted at Step 2 to transform the training set S into a collection of binary strings that can be viewed as a portion of the truth table of a monotone Boolean function. Here, for each (binarized version of a) pattern $\boldsymbol{x}$ in S, the output is the class y, possibly coded in binary form if there are more than two classes.

To ensure a good generalization ability, the logic synthesis (Step 3) is performed via an optimized version of the Shadow Clustering (SC) algorithm,[16] a proper technique for reconstructing monotone Boolean functions starting from a partially defined truth table. In contrast with methods based on a divide-and-conquer approach, SC adopts an aggregative policy, that is, at any iteration some patterns (coded in binary form) belonging to the same output class are clustered to produce an intelligible rule. A suitable heuristic approach is employed to generate implicants (rules) exhibiting the highest covering and the lowest error; a trade-off between these two different objectives generally leads to final models showing a good accuracy.

The training algorithm for LLM requires to define a single parameter ε, the maximum error that can be scored by each generated rule. In all our trials, we have used the value $\varepsilon=0$.

Results

Table 2 resumes the performance of standard clinical classification in leave-one-out cross-validation, based on the IPS index, and that of the selected supervised methods. IPS correctly classified 68 per cent of total patients, with 37 per cent sensitivity and 80 per cent specificity.

Among the three black-box methods, the best performance was achieved by SVM with linear kernel (global accuracy = 82%, sensitivity =55%, specificity = 92%). kNN with $k=1$ also outperformed the standard clinical classification (global accuracy = 75%, sensitivity = 45%, specificity = 87%), whereas models with higher k values showed a poor performance and, in particular, a very low sensitivity. With regard to ANN, the model with two hidden neurons shows the highest performance, which lies between that of IPS only and that of kNN (global accuracy = 72%, sensitivity = 45%, specificity= 83%).

Among the two considered rule-based methods, LLM showed the best performance (global accuracy =70% vs 65% for DT), even if sensitivity was slightly lower (45% vs. 47%).

When the analysis was repeated on the whole data set, LLM selected 25 rules that included a minimum of two and a maximum of six conditions; the corresponding covering ranged between 2.2 and 53.3 per cent.

Table 3 shows the rules generated by LLM after the exclusion of those with a low coverage (<20%). This restriction was made in order to reduce the effect of outliers, thus allowing a more reliable comparison with DT after the pruning procedure.

All the LLM rules included at least one clinical or demographic characteristic of patients. On the whole, LLM identified four features relevant for classification, all inversely associated with the occurrence of relapse (namely, *MS4A3, RPS8, DMD* and *MUC5AC*). With regard to clinical conditions, advanced stages (3 and 4) were more often associated with relapse, but with some exceptions (e.g. rule 2, condition 1). IPS was included in only one rule (no. 6), and as expected, a low value corresponded to the absence of relapse. Finally, gender was included in 6 out of 13 rules. Among the four rules identifying relapsed patients, two included males (no. 10 and no. 12, respectively, condition 1), whereas females were never selected.

Figure 1 shows the classifier obtained by DT. Classification was performed by seven rules that involved gene expression only (namely, *XIST, EPOR, GPR82, AV719529* and *KIAA1430*). A prediction of relapse was associated with low values of *XIST* and *GPR82* and high values of *AV19529* and *KIAA1430*.

Discussion

Despite advances in therapeutic treatment, about 20 per cent of HL patients eventually die, whereas a similar proportion is likely to be over-treated.[8] The large availability of new potential tumour markers for HL prognosis, including genome-wide gene expression data, might contribute to the improvement of the performance of IPS in predicting patient survival.[7,8,19]

Table 2. Comparison between standard clinical classification by IPS score and the selected methods of supervised analysis in leave-one-out cross-validation.

Classification method	Global accuracy		Sensitivity		Specificity	
	N	%	*N*	%	*N*	%
Standard clinical classification						
IPS	88	67.7	14	36.8	74	80.4
Black-box methods						
kNN						
k = 1	97	74.6	17	44.7	80	87.0
k = 3	84	64.6	7	18.4	77	83.7
k = 5	90	69.2	6	15.8	84	91.3
ANN						
One hidden neuron	92	70.8	17	44.7	75	81.5
Two hidden neurons	93	71.5	17	44.7	76	82.6
Three hidden neurons	91	70.0	13	34.2	78	84.8
SVM						
RBF kernel	92	70.8	0	0.0	92	100
Linear kernel	106	81.5	21	55.3	85	92.4
Rule-based methods						
DT	85	65.4	18	47.4	67	72.8
LLM	91	70.0	17	44.7	74	80.4

kNN: *k*-Nearest Neighbour classifier; ANN: Artificial Neural Network; SVM: Support Vector Machine; RBF: Radial Basis Function; LLM: Logic Learning Machine; DT: Decision Tree; IPS: International Prognostic Score; *N*: number of patients correctly classified.

Table 3. Classification rules identified by the Logic Learning Machine on the whole data set.

No.	Relapse	Condition 1	Condition 2	Condition 3	Condition 4	Covering %
1	No	Stage 1 or 2	MS4A3 > 1.729	MUC5AC > 2.701	–	53.3
2	No	Stage 3	MS4A3 > 1.729	RPS8 > 11.432	–	51.1
3	No	Stage 2 or 4	RPS8 > 11.432	MUC5AC > 2.701	–	40.2
4	No	Female gender	MS4A3 > 1.729	–	–	38.0
5	No	Stage 1 or 2	RPS8 > 11.432	DMD > 1.989	–	34.8
6	No	MS4A3 > 1.729	DMD > 1.989	MUC5AC > 2.701	Low IPS	34.8
7	No	Male gender	Stage 2	MUC5AC > 2.701	–	29.3
8	No	Female gender	Stage 3	DMD > 1.989	–	27.2
9	No	Male gender	Stage 3	RPS8 > 11.432	MUC5AC > 2.701	26.1
10	Yes	Male gender	Stage 3 or 4	MUC5AC ⩽ 2.701	–	26.3
11	Yes	Stage 2 or 4	MS4A3 ⩽ 1.729	MUC5AC ⩽ 2.701	–	26.3
12	Yes	Male gender	Stage 1 or 4	RPS8 ⩽ 11.432	–	21.1
13	Yes	Stage 3	RPS8 ⩽ 11.432	DMD ⩽ 1.989	MUC5AC ⩽ 2.701	21.1

IPS: International Prognostic Score.
The 13 out of 25 rules with at least 20 per cent of covering are shown.

Many supervised methods of data analysis are available to exploit and combine information from new tumour markers and clinical prognostic factors. In particular, ANN, kNN and the more recent SVM have shown a high accuracy in predicting survival of cancer patients when applied to

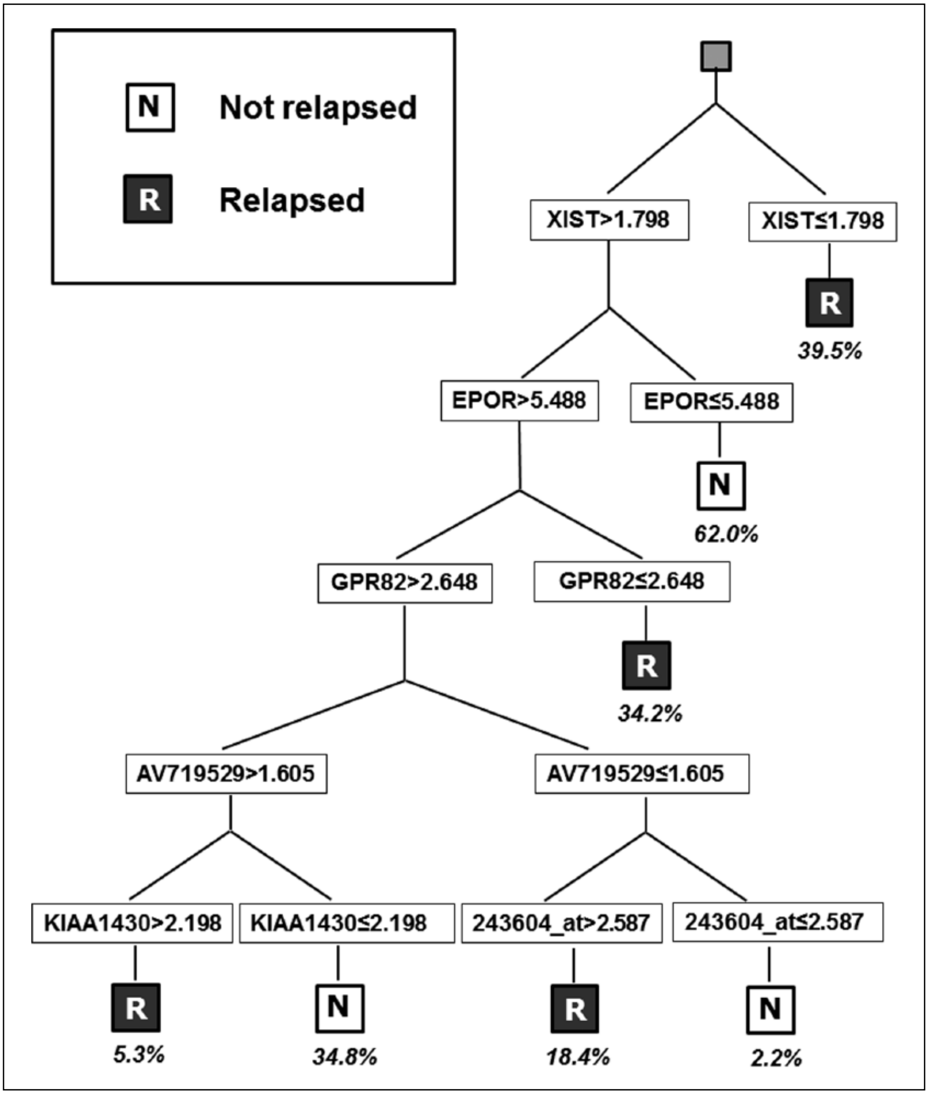


Figure 1. Classification obtained by DT on the whole data set. Percentages indicate the covering of each rule.

gene expression data in many different clinical settings.[20–26] However, such algorithms are usually referred to as 'black-box' methods since classification is made through a mathematical formula that makes it difficult to evaluate the biological and clinical role of variables included in the analysis. Conversely, algorithms based on intelligible threshold rules, like DT and the recently proposed LLM, can provide useful information for a better understanding of tumour biology and for addressing therapeutic approaches.[18,23]

The good performance of LLM compared to that of common supervised techniques was demonstrated in a set of biomedical studies.[18,27,28] However, different from DT,[20–29] LLM has been never applied for classification purposes to large databases of highly correlated features, such as microarray gene expression data. In this study, in agreement with results from previous investigations, LLM showed a performance quite similar to that of some common competing black-box methods (ANN and kNN), but lower than that of SVM.

LLM outperformed DT and was able to combine information from clinical variables with expression values from a small panel of selected genes. In particular, stage and gender were in some cases associated in the same rule, but never associated with IPS (Table 2). Since IPS is constructed using clinical variables that also include stage and gender,[5] this finding suggests that LLM tends to reject redundant information. Furthermore, a low IPS score, a low stage at diagnosis and female gender were more often associated with a good prognosis, in agreement with knowledge from previous investigations.[5]

Taken together, these results suggest that the combination of clinical data and gene expression features could provide useful information for assessing the prognosis of HL patients. This observation is in agreement with previous studies on different malignancies, indicating that clinical information can enrich microarray data in identifying a suitable classifier for the prediction of cancer survivability.[23,30,31]

Gene expressions selected by LLM were all different from those identified by DT, and they also differed from the 30 most relevant features identified by the original analysis. However, *MUC5AC* and *EPOR* were also included into the complete list of differentially expressed genes reported by Steidl et al.[8] The four genes identified by LLM were all under-expressed in relapsed patients. *MS4A3* (membrane-spanning 4-domains subfamily A member 3) is localized in 11q12 and encodes a membrane protein probably involved in signal transduction.[32] Interestingly, *MS4A3* belongs to the same membrane-spanning 4-domains gene subfamily of *MS4A4*, which was recognized to be associated with HL prognosis in previous investigations.[33] *RPS8* is localized in 1p34.1-p32 and encodes a ribosomal protein that is a component of the 40S subunit.[34] *DMD* (dystrophin) locates at Xp21.2 and is a highly complex gene, containing at least eight independent, tissue-specific promoters and two polyA-addition sites.[35] Finally, *MUC5AC* is located in 11p15.5[36] and encodes for a protein (mucin) involved in secretion of gastrointestinal mucosa.

With regard to DT, genes with a known function (http://www.ncbi.nlm.nih.gov/gene) include the following: *XIST* (X inactive specific transcript), which is a non-coding gene located in Xq13.2, involved in the inactivation of X chromosome in human females,[37] and *EPOR*, located in 19p13.3-p13.2, which encodes an erythropoietin receptor.[38] Moreover, *GPR82*, localized in Xp11.4, encodes for a protein with unknown function but is a member of a family of proteins that contain seven transmembrane domains and transduce extracellular signals through heterotrimeric G proteins.[39] Interestingly, partly consistently with our observation of a higher relapse probability among subjects with low *XIST* expression, *XIST* was demonstrated to activate apoptosis in T lymphoma cells via ectopic inactivation of the X chromosome.[40] In our data, *XIST* was strongly overexpressed among females (data not shown), thus potentially providing a new insight about the biological mechanism at the basis of the better prognosis commonly observed among females.

Results of our study may be prone to some limitations. In particular, we selected the GDS4222 data set because, at least to our knowledge, it was among the biggest publicly available gene expression databases including information about prognosis of HL patients. However, as a whole, its sample size (130 patients, including 38 relapsed) was too small to allow drawing definitive conclusions, and all findings reported in our study need confirmation by other independent investigations. Furthermore, sensitivity of any applied method (including SVM) was unsatisfactory (<60%). In fact, in the presence of unbalanced outcomes, as in our study, rules extracted from LLM can be weighted to improve their accuracy.[17] According to this property, we tried to reclassify a posteriori the patients under study by assigning a 1:10,000 weight in favour of relapsed outcome, but also in this further analysis sensitivity never achieved 60 per cent (data not shown), pointing out that the limit of 60 per cent for sensitivity is difficult to be exceeded for any of the considered methods. The lack of potentially relevant clinical information (e.g. absolute lymphocyte count, age at diagnosis and first line treatment) and the poor measure of the outcome, which did not include time-to-event values, could have contributed to lowering the sensitivity of our study. Moreover, we performed all the analyses without applying any pre-filtering technique to the data under study. Previous investigations have demonstrated that the performance of supervised methods can be enhanced by applying pre-filtering and feature selection methods, which can reduce overfitting.[41–43] Their effect on LLM classification has not been investigated yet.

Conclusion

LLM provided simple intelligible rules that could contribute to the knowledge of HL biology and to address therapeutic approaches by combining clinical information and gene expression data.

The role of genes identified by both LLM and DT in the clinical course of HL patients should be investigated in further studies. In particular, the higher expression of *XIST* in patients with a good outcome and among females might be related to the still unknown factors favouring the better prognosis of female patients with HL.

Acknowledgements

Stefano Parodi is a research fellow of the Italian MIUR Flagship project ‘InterOmics’.

Declaration of conflicting interests

The author(s) declared no potential conflicts of interest with respect to the research, authorship and/or publication of this article.

Funding

The author(s) received no financial support for the research, authorship and/or publication of this article.

References

1. Parodi S and Stagnaro E. *Hodgkin’s Disease Worldwide – Incidence, Mortality, Survival, Prevalence and Time Trend*. New York: Nova Science Publisher, 2009, pp. 1–8.
2. Banerjee D. Recent advances in the pathobiology of Hodgkin’s Lymphoma: potential impact on diagnostic, predictive, and therapeutic strategies. *Adv Hematol* 2011; 2011: 439456.
3. Swerdlow SH, Campo E, Harris NL, et al. *WHO Classification of Tumours of Haematopoietic and Lymphoid Tissues*. 4th ed. Lyon: IARC Press, 2008.
4. Ansell SM. Hodgkin Lymphoma: 2014 update on diagnosis, risk-stratification, and management. *Am J Hematol* 2014; 89: 771–779.
5. Hasenclever D and Diehl V. A prognostic score for advanced Hodgkin’s disease: international prognostic factors project on advanced Hodgkin’s disease. *New Engl J Med* 1998; 339: 1506–1514.
6. King RL, Howard MT and Bagg A. Hodgkin Lymphoma: pathology, pathogenesis, and a plethora of potential prognostic predictors. *Adv Anat Pathol* 2014; 21: 12–25.
7. Cuccaro A, Bartolomei F, Cupelli E, et al. Prognostic factors in Hodgkin lymphoma. *Mediterr J Hematol Infect Dis* 2014; 6: e2014053.
8. Steidl C, Lee T, Shah SP, et al. Tumor-associated macrophages and survival in classic Hodgkin’s lymphoma. *New Engl J Med* 2010; 362: 875–885.
9. Azambuja D, Natkunam Y, Biasoli I, et al. Lack of association of tumor-associated macrophages with clinical outcome in patients with classical Hodgkin’s lymphoma. *Ann Oncol* 2012; 23: 736–742.

10. Sánchez-Espiridión B, Martin-Moreno AM, Montalbán C, et al. Immunohistochemical markers for tumor associated macrophages and survival in advanced classical Hodgkin's lymphoma. *Haematologica* 2012; 97: 1080–1084.
11. Edgar R, Domrachev M and Lash AE. Gene expression omnibus: NCBI gene expression and hybridization array data repository. *Nucleic Acids Res* 2002; 30: 207–210.
12. Dudoit S and Fridlyand J. Classification in microarray experiments. In: Speed T (ed.) *Statistical analysis of gene expression microarray data*. Boca Raton, FL: Chapman & Hall, 2003, pp. 93–158.
13. Michie D, Spiegelhalter D and Taylor C. *Machine learning: neural and statistical classification*. Chichester: Ellis Horwood, 1999.
14. Quinlan JR. *C4.5 programs for machine learning*. San Francisco, CA: Morgan Kaufmann Publishers, 1992.
15. Muselli M. Switching neural networks: a new connectionist model for classification. In: Apolloni B, Marinaro M, Nicosia G, et al. (eds) *Lecture notes in computer science* (vol. 3931). Berlin: Springer-Verlag, 2006, pp. 23–30.
16. Muselli M and Ferrari E. Coupling logical analysis of data and shadow clustering for partially defined positive Boolean function reconstruction. *IEEE T Knowl Data En* 2011; 23: 37–50.
17. Ferrari E and Muselli M. Maximizing pattern separation in discretizing continuous features for classification purposes. In: *Proceeding of the 2010 international joint conference on neural networks (IJCNN)*, Barcelona, 18–23 July 2010.
18. Cangelosi D, Muselli M, Parodi S, et al. Use of attribute driven incremental discretization and logic learning machine to build a prognostic classifier for neuroblastoma patients. *BMC Bioinformatics* 2014; 15(Suppl. 5): S4.
19. Montalbán C, García JF, Abraira V, et al. Influence of biologic markers on the outcome of Hodgkin's lymphoma: a study by the Spanish Hodgkin's lymphoma study group. *J Clin Oncol* 2004; 22: 1664–1673.
20. Chen HY, Yu SL, Chen CH, et al. A five-gene signature and clinical outcome in non-small-cell lung cancer. *New Engl J Med* 2007; 356: 11–20.
21. Chen YC, Ke WC and Chiu HW. Risk classification of cancer survival using ANN with gene expression data from multiple laboratories. *Comput Biol Med* 2014; 48: 1–7.
22. Shi M, Beauchamp RD and Zhang B. A network-based gene expression signature informs prognosis and treatment for colorectal cancer patients. *PLoS One* 2012; 7: e41292.
23. Cruz JA and Wishart DS. Applications of machine learning in cancer prediction and prognosis. *Cancer Inf* 2007; 2: 59–77.
24. Sørlie T, Perou CM, Fan C, et al. Gene expression profiles do not consistently predict the clinical treatment response in locally advanced breast cancer. *Mol Cancer Ther* 2006; 5: 2914–2918.
25. Lisboa PJ and Taktak AF. The use of artificial neural networks in decision support in cancer: a systematic review. *Neural Netw* 2006; 19: 408–415.
26. Barrier A, Lemoine A, Boelle PY, et al. Colon cancer prognosis prediction by gene expression profiling. *Oncogene* 2005; 24: 6155–6164.
27. Muselli M, Costacurta M and Ruffino F. Evaluating switching neural networks through artificial and real gene expression data. *Artif Intell Med* 2009; 45: 163–171.
28. Muselli M. Extracting knowledge from biomedical data through Logic Learning Machines and Rulex. *EMBnet J* 2012; 18: 56–58.
29. Irshad S, Bansal M, Castillo-Martin M, et al. A molecular signature predictive of indolent prostate cancer. *Sci Transl Med* 2013; 5: 202ra122.
30. Futschik ME, Reeve A and Kasabov N. Evolving connectionist systems for knowledge discovery from gene expression data of cancer tissue. *Artif Intell Med* 2003; 28: 165–189.
31. Miyake H and Fujisawa M. Prognostic prediction following radical prostatectomy for prostate cancer using conventional as well as molecular biological approaches. *Int J Urol* 2013; 20: 301–311.
32. Adra CN, Lelias JM, Kobayashi H, et al. Cloning of the cDNA for a hematopoietic cell-specific protein related to CD20 and the beta subunit of the high-affinity IgE receptor: evidence for a family of proteins with four membrane-spanning regions. *Proc Natl Acad Sci U S A* 1994; 91: 10178–10182.

33. Steidl C, Connors JM and Gascoyne RD. Molecular pathogenesis of Hodgkin's lymphoma: increasing evidence of the importance of the microenvironment. *J Clin Oncol* 2011; 29: 1812–1826.
34. Davies B and Fried M. The structure of the human intron-containing S8 ribosomal protein gene and determination of its chromosomal location at 1p32-p34.1. *Genomics* 1993; 15: 68–75.
35. Zimowski J, Fidziańska E, Holding M, et al. Two mutations in one dystrophin gene. *Neurol Neurochir Pol* 2013; 47: 131–137.
36. Guyonnet Duperat V, Audie JP, Debailleul V, et al. Characterization of the human mucin gene MUC5AC: a consensus cysteine-rich domain for 11p15 mucin genes? *Biochem J* 1995; 305: 211–219.
37. Weakley SM, Wang H, Yao Q, et al. Expression and function of a large non-coding RNA gene XIST in human cancer. *World J Surg* 2011; 35: 1751–1756.
38. Lisowska KA, Frackowiak JE, Mikosik A, et al. Changes in the expression of transcription factors involved in modulating the expression of EPO-R in activated human CD4-positive lymphocytes. *PLoS One* 2013; 8: e60326.
39. Lee DK, Nguyen T, Lynch KR, et al. Discovery and mapping of ten novel G protein-coupled receptor genes. *Gene* 2011; 275: 83–91.
40. Agrelo R, Souabni A, Novatchkova M, et al. SATB1 defines the developmental context for gene silencing by Xist in lymphoma and embryonic cells. *Dev Cell* 2009; 16: 507–516.
41. Bala J, Huang J, Vafaie H, et al. Hybrid learning using genetic algorithms and decision trees for pattern classification. In: Proceedings of the 14th international joint conference on Artificial intelligence, Montreal, WI, 20 August 1995, pp. 719–724. San Francisco, CA: Morgan Kaufmann Publishers.
42. Hsu WH. Genetic wrappers for feature selection in decision tree induction and variable ordering in Bayesian network structure learning. *Inform Sciences* 2004; 163: 103–122.
43. Hajiloo M, Rabiee HR and Anooshahpour M. Fuzzy support vector machine: an efficient rule-based classification technique for microarrays. *BMC Bioinformatics* 2013; 14(Suppl. 13): S4.

Health Informatics Journal

Article

Health Informatics Journal
2018, Vol. 24(1) 66–91
© The Author(s) 2016
Reprints and permissions:
sagepub.co.uk/journalsPermissions.nav
DOI: 10.1177/1460458216650979
journals.sagepub.com/home/jhi
SAGE

Persuasive technology for health and wellness: State-of-the-art and emerging trends

Rita Orji and Karyn Moffatt
McGill University, Canada

Abstract

The evolving field of persuasive and behavior change technology is increasingly targeted at influencing behavior in the area of health and wellness. This paper provides an empirical review of 16 years (85 papers) of literature on persuasive technology for health and wellness to: (1.) answer important questions regarding the effectiveness of persuasive technology for health and wellness, (2.) summarize and highlight trends in the technology design, research methods, motivational strategies, theories, and health behaviors targeted by research to date, (3.) uncover pitfalls of existing persuasive technological interventions for health and wellness, and (4.) suggest directions for future research.

Keywords

behavior change systems, behavior theories, gamification, health and wellness, health interventions, health technology, healthy behavior, motivational strategies, persuasive strategies, persuasive technology

Introduction

Persuasive Technology (PT) are interactive systems designed to aid and motivate people to adopt behaviors that are beneficial to them and their community while avoiding harmful ones. The use of PT, aimed at bringing about desirable change by shaping and reinforcing behavior and/or attitude is growing in virtually all areas of health and wellness. Over the past decade, several PT have been developed targeted at impacting one or more aspects of health and wellness. These technologies can broadly be classified into two main category: PT for health promotion and prevention and PT for disease management.[1,2] PT for health promotion and prevention are targeted at behaviors undertaken by individuals for the purposes of preventing illness, detecting early illness symptoms, and maintaining general wellbeing.[3] Examples include being physical activity,[4–6] healthy eating,[7–9] smoking cessation,[10,11] avoiding risky sexual behavior and unwanted pregnancy,[12,13] and dental health.[14–16] PT for disease management help patients improve health-related self-management

Corresponding author:
Rita Orji, McGill University, 3661 Peel Street, Montreal, QC H3A 0G4, Canada.
Email: rita.orji@mail.mcgill.ca

skills such as teaching them how to manage certain illnesses, helping them comply and adhere to treatment directives.[17] Each of these health behavior domains has attracted considerable attention.

There is an increasing interest and investments in developing and using technology to promote health and wellness by various stakeholders including health and wellness researchers and practitioners, technology designers, and public health and government agencies. Therefore, it is necessary to conduct an empirical review to reevaluate and uncover important trends, best practices, gaps, and opportunities for improvement. In addition, research on this topic is fragmented, using many different approaches, methods and concepts. A literature review can help bring these disparate sources together.

Thus, in this paper, we present an empirical review of 16-years (from 2000 to 2015) of PT studies across various health and wellness domains with the aim of: (1.) answering important questions regarding the effectiveness of persuasive technology for health and wellness; (2.) highlighting and summarizing emerging trends in the technological intervention design, research method, target health behavior, use of motivational strategies and behavior theories – which is important in guiding and setting roadmap for subsequent research agenda; (3.) uncovering pitfalls of existing PT interventions for health; and finally, (4.) suggesting directions for future research. This review serves as a reference for future research in this area, providing a comprehensive overview that will be a useful starting point for anyone interested in an overview of persuasive technology for health and wellness by systematically analyzing and categorizing the scattered research effort in this area under useful headings and highlighting the merging trends.

Materials and methods

As our goal is to systematically analyze persuasive technology in the health domain, we employed quantitative content analysis, a technique which enables comparison, contrast, and categorization of data according to different themes and concepts.[18] This entails collecting data in a rigorous way, paying special attention to the objectivity of the results.

For our literature search, we used the Elsevier Scopus database as our first data source with the search terms "Persuasive Health Technology", "Persuasive Technology and Health", "Behavior Change Technology and Health" "Persuasive Technology", "Technology and Health Interventions". Scopus is the largest abstract and citation database of peer-reviewed literature.[19] We also searched PubMed, EBSCOHost, Springer, the ACM Digital Library, IEEE Xplore, and Google Scholar with the same search term. This ensures good coverage of technological health interventions across various fields including Human-Computer Interaction (HCI), medical and health informatics, health information systems, and other related research field. Finally, we scanned through the reference lists of the included studies to find further potentially relevant studies. The search resulted in 1842 unique titles, of which 544 articles were deemed relevant following a title examination. After the abstracts of each article were reviewed, a total of 85 articles that were published from 2000 to 2015 are included in this analysis. We included only articles that discussed the design and evaluation of new PT for health and wellness or an evaluation of existing PT for health and wellness and are published in English. We also excluded papers describing the design and development of PT for health without an evaluation. The study identification process is as summarized in Figure 1.

Analysis and coding scheme

In the second stage of the review, we coded the articles. To achieve this, we iteratively developed a coding sheet for analyzing PT, see Table 1. Next, we went through each of the articles and classified their data using the coding sheet. The coding sheet included the following parts (see Table 1):

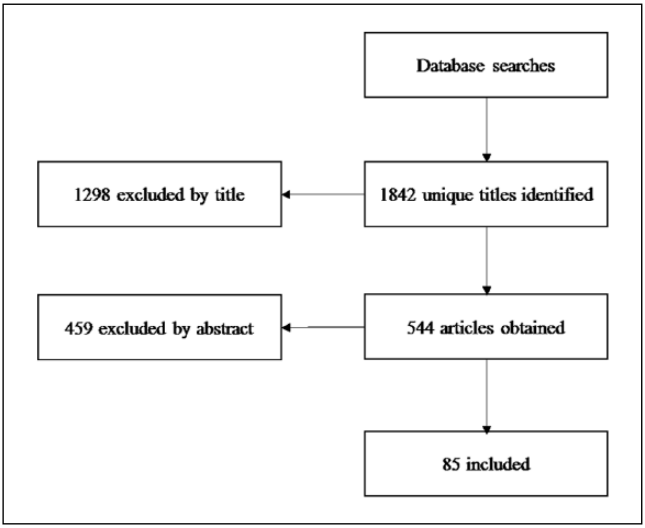


Figure 1. Included study identification process.

Table 1. Persuasive technology classification and analysis coding scheme.

S/N	Identification	Name, author(s), year, publication venue
1	Targeted (Health) Domain	Physical activity, Eating, smoking, etc.
2	Technology	Web, mobile, games, desktop applications, etc.
3	Duration of Evaluation	Hours, Days, Weeks, Months, and Years.
4	Behavior Theories	Theories Employed in the PT design or evaluation
5	Motivational Strategies	Motivational affordance employed in PT design
6	Evaluation Method	Quantitative, Qualitative, and Mixed.
7	Targeted Age group	Children, Adults, Elderly, etc.
8	Number of Participants	Number of participants involved in the evaluation.
9	Study Country	Country where the study was conducted.
10	Targeted behavioral or psychological outcome	Behavior, Attitude, Adherence, etc.
11	Findings/Results	Whether successful or not.

Results

The analysis of existing PT for health and wellness revealed some interesting insights and trends. Below we present our findings under various categories including: evaluation outcome, employed technology platform, persuasive and motivational strategies, behavior theories, targeted behavior domain, theory, strategy, and outcome mapping. The detailed summary of all the reviewed studies is presented in the Appendix.

Persuasive health technology by year and country

As shown in Figure 2, a relatively large proportion of empirical studies of PT for health and wellness are published after 2005 compared to before 2005. However, there was a big jump starting in

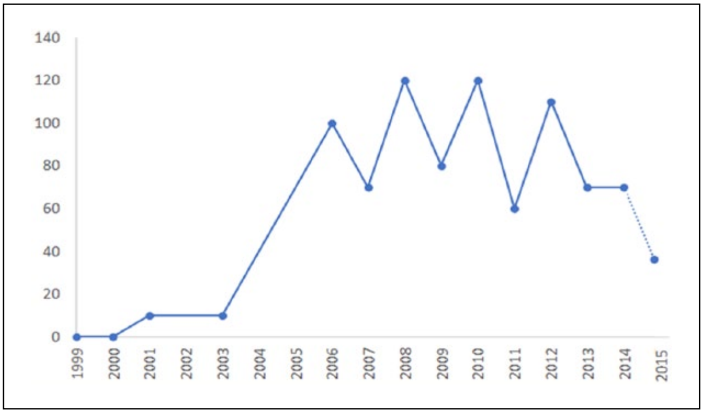


Figure 2. Persuasive technology for health and wellness trend by year.

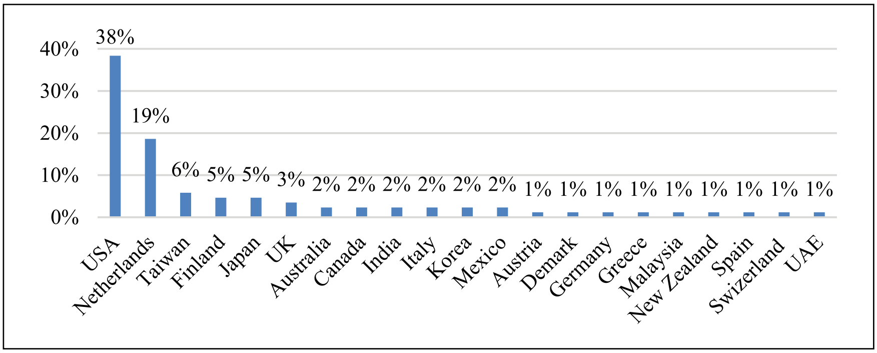


Figure 3. Persuasive technology for health and wellness trend by study country.

2005, and after that it has been relatively stable, despite some year to year fluctuation. It is important to highlight that although the year 2015 seems to have the least number of studies since 2005 (Figure 2), it is probably because the study was completed halfway through 2015, with many of the publications for the year still pending.

As shown in Figure 3, the studies were conducted in 21 different countries with the USA leading the list with a total of 38% of all the studies. USA is followed by the Netherlands with 19%. Taiwan is in the third place with a total of 6% and Finland and Japan are the fifth place, having 5% each of all the studies.

Evaluation outcome: does persuasive health technology work? Effectiveness of persuasive health technology

Figure 4, summarizes the reported results from the evaluation of the PT for health and wellness reviewed in this paper. Of the 85 reviewed studies, 64 (75%) reported fully positive outcome from using the PT to impact specified health behavior. Fourteen studies (17%) reported results partially positive – a combination of positive with negative or no effect results. Only 7 (8%) of all the studies were unsuccessful at achieving their intended persuasion objective – finding negative results,

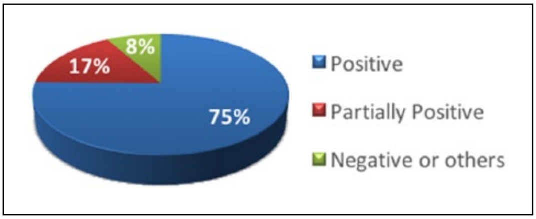


Figure 4. Summary results of the effectiveness of persuasive technology.

Table 2. Detailed results of the effectiveness of persuasive technology.

Outcome	Study	Total	% overall
Positive	[5]–[9], [12]–[17], [21]–[70]	64	75%
Partially Positive	[10], [11], [71]–[82]	14	17%
Negative or others	[15], [83]–[88]	7	8%

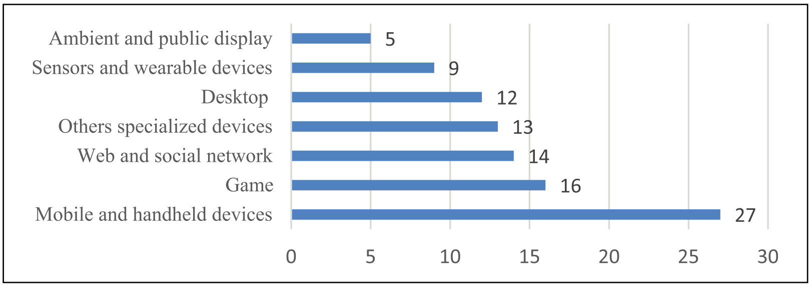


Figure 5. Persuasive technology platforms.

no positive results, or no results at all.[20] The results of the effectiveness of PT for health and wellness are detailed in Table 2.

Technology platforms for persuasive technology

Figure 5 summarizes the major technology platforms employed by PT for health and wellness designers. The most frequently employed technology platforms emerged to be mobile and handheld devices with a total of 27 (28%). It is followed by games with 16 (17%). Games category included all the studies that delivered their PT interventions in the form of games irrespective of whether the game is mobile-based, web, or runs on a stand-alone desktop. In addition, persuasive implementations on the web and social networking sites are common among the studies reviewed. Ambient and public display is the least frequently employed platform with 5(5%).

Persuasive strategies and motivational affordance employed

The reviewed studies employed several persuasive strategies and motivational affordances to bring about the intended persuasion outcomes. Table 3 shows the prevalent motivational affordance and

Table 3. Persuasive strategies/motivational affordances by Persuasion Outcome.

Motivational strategies/ affordances	Studies with positive result	Studies with partially positive result	Studies with negative result or others	Total
Tracking and monitoring	[4], [6], [8], [14], [21], [22], [27], [29], [33]–[38], [46]–[50], [55], [56], [58], [60], [62], [67], [70], [89]	[10], [73], [75], [79]	[15], [85], [86]	34
Audio, Visual and Textual Feedback	[4], [6], [7], [9], [14], [16], [27], [29], [30], [37], [43], [46], [48], [50], [65]–[67], [69]	[11], [71], [72]	[15], [83]–[85], [87], [88]	28
Social support, sharing, and comparison	[22], [24], [25], [27], [30], [41], [32], [35], [44], [45], [46], [48], [53], [65], [66], [4], [21]	[75]	[84], [88]	23
Persuasive messages, reminder, and alert	[22], [30], [34], [35], [37], [42], [49], [53], [55], [58], [60], [63]	[72]–[75], [77]	[86], [87]	19
Reward, points, credits	[6]–[9], [17], [25], [27], [30], [34], [38], [39], [47], [59], [61], [69]	[76]	[15], [86]	17
Goal and Objectives	[4], [6], [9], [22], [25], [46], [47], [57]	[75], [77], [79]	[86]	13
Competition, leaderboards, ranking	[4], [17], [22], [34], [39], [41], [47], [52], [56], [62]		[88]	11
Tailoring, Personalization and customization	[12], [17], [38], [40], [68], [70]	[72]	[86]	8
Praise	[14], [22], [24], [40]	[78]	[83], [86]	7
Cooperation and Collaboration	[4], [32], [47], [52]	[81]		5
Virtual rehearsal and Simulation	[12], [24], [63], [69]			4
Emoticons and persuasive images	[15], [26], [33]			3
Progress	[4], [6]		[87]	3
Positive Reinforcement	[6], [15]	[78]	[87]	3
Negative Reinforcement	[8]	[78]	[87]	3
Suggestions and advice	[5], [14]	[72]		3
Video-based persuasion	[23]			1
Not Specified	[13], [28], [31], [51], [54], [64]	[80], [82]		8

persuasion outcome. Tracking and monitoring is the most frequently employed strategy (with a total of 34 studies), followed by feedback (28). It is important to note that approximately 80% of the reviewed studies employed more than one motivational strategy (see the Appendix) and are categorized accordingly. Some of the studies did not specify their strategy.

Behavior theories employed

Examining behavior change theories employed reveals that more than half of all the studies reviewed (55%) are not informed by any theory or did not specify the theories that inform their PT

Table 4. Behavior theories employed in persuasive technology design.

Theories	Study	Total
Transtheoretical Model	[4], [6], [8], [9], [11], [25], [27], [30], [33], [57], [64], [72], [89]	13
Goal Setting Theory	[15], [23], [75], [77], [79]	5
Social Conformity Theory	[35], [57], [69]	3
Theory of Reasoned Action	[21], [69]	2
Self Determination Theory	[29], [54]	2
Unified Theory of Acceptance and Use of Technology	[26], [80]	2
Reinforcement Theory	[15], [38]	2
Social Cognitive Theory	[23]	1
Ego Depletion theory	[31]	1
Premack's principle	[38]	1
Parallel Process Model	[13]	1
Theory of Meaning Behavior	[54]	1
Sexual Health Model	[12]	1
Social Learning Theory	[78]	1
Health Belief Model	[69]	1
Theory of Planned Behavior	[44]	1
Big Five Personality Theory	[80]	1
Knowledge, Attitude, Behavior Model	[9]	1
Cognitive Behavior Therapy	[57]	1
Technology Acceptance Model	[30]	1
Not specified	[5], [7], [10], [16], [17], [22], [24], [28], [32], [34], [36], [37], [39]–[43], [45]–[53], [55], [56], [58]–[63], [65]–[68], [70], [71], [73], [74], [76], [81]–[88]	51

intervention design, see Table 4. Even among the studies that specified the theories that informed their design, most of them only mentioned the theories without actually specifying how the theories informed the actual PT intervention design components and/or evaluation. Transtheoretical model of change (TTM) is the most frequently employed theory with a total of 13 (14%) studies. Most of the studies based on theories employed more than one theory or adapted constructs from more than one theory.

Targeted health behavior domain of persuasive technology by persuasion outcome, motivational strategies, and behavior theories

As can be seen from Figure 5 and Table 5, the PT for health and wellness reviewed in this paper fundamentally focused at imparting eight major health behavior domain, including *physical activity, eating, dental health, disease management, smoking and substance use, sexual behavior, general health*, and *others*. "Others" consists of health behaviors that appeared less frequently (in all cases, only one study looked to the behavior) such as sleeping, and depression. Physical activity has attracted the most research interest, making up 38% of all the reviewed studies, followed by healthy eating with a total of 25%.

Table 5. Targeted Health Domains of Persuasive Technologies by Persuasion Outcome, Motivational Strategies, and Behavior Theories – only frequently employed strategies and theory are reported for domains with many strategies such as physical activity and eating.

Health Domain	Studies with positive result	Studies with partially positive result	Studies with negative result or others	% of all	Motivational Strategies (no. of PTs employing each strategy)	Behavior theories (no. of PTs employing each theory)
Physical Activity	[4]–[6], [21], [22], [25], [27], [29], [32], [38]–[41], [46]–[48], [50], [52], [54], [56], [59], [61], [62], [66]	[73]–[75], [79]	[15], [85], [86], [88]	38%	Tracking and monitoring (19) Social support, sharing, and comparison (17) Competition, leaderboard, ranking (10) Reward, points, credits (10) Goal and Objectives (9) Audio, Visual and Textual Feedback (8)	Transtheoretical Model (4) Goal Setting Theory (3) Theory of Meanings of Behavior (2) Self-determination theory (2) Operant Conditioning (2)
Eating	[7]–[9], [17], [28], [30], [31], [33], [35], [45], [55], [60], [65], [67]–[70], [89]	[78]	[83]	25%	Audio, Visual and Textual Feedback (8) Tracking and monitoring (7) Social support, sharing, and comparison (6) Reward, points, credits (5)	Transtheoretical Model (4) Health belief model Social learning theory Social Cognitive Theory Knowlegde-Attitide-Behavior Model
Dental Health	[14]–[16], [24]			5%	Audio, Visual and Textual Feedback (3) Tracking and monitoring (2)	Operant Conditioning (2)
Disease Management	[36], [58]	[77], [80]	[84]	6%	Tracking and monitoring Goal and Objectives Persuasive messages, reminder, and alert	Goal Setting Theory
Smoking and Substance Abuse		[10], [11], [72]		4%	Social support, sharing, and comparison Goal and Objectives Tracking and monitoring	Transtheoretical Model(2)
Sexual behavior (HIV and STD)	[12], [13]			2%	Virtual rehearsal and Simulation	Parallel Process Model Sexual Health Model
General Health	[51], [53]	[82]		4%	Social support, sharing, and comparison Persuasive messages, reminder, and alert	None
Others	[23], [26], [34], [37], [42]–[44], [49], [57], [63], [64]	[71], [76], [81]	[87]	18%	Progress Positive and negative Reinforcement Suggestion Praise	Cognitive-Behavior Therapy Acceptance and Commitment Therapy Big Five Personality Theory Technology Acceptance Model Premack's principle Ego Depletion theory Reinforcement Theory Unified Theory of Acceptance and Use of Technology Theory of Reasoned Action Social Conformity Theory

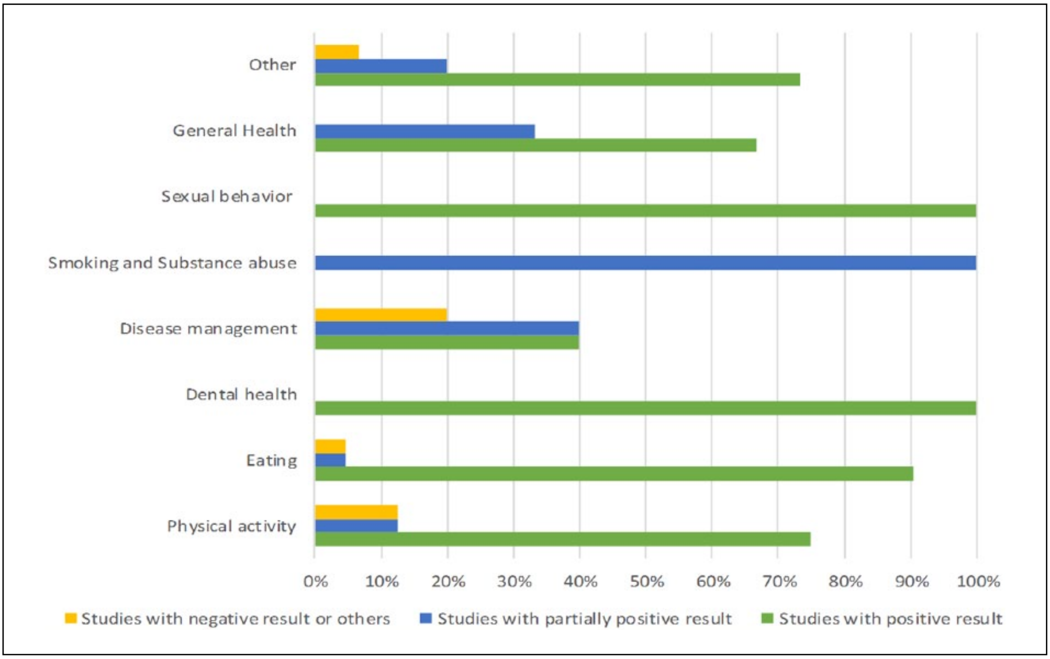


Figure 6. Comparative effectiveness of PT by health domain.

Health behavior domain and persuasion outcome. With respect to the persuasion outcomes, it is difficult to speak specifically to the domain specific effectiveness of PT for health and wellness because of the variability in the number of studies from each domain included in this analysis and other domain-specific factors that could influence the effectiveness of PT including the instantiation and operationalization of the persuasive strategies and the duration of evaluation. However, based on the results from the analyzed studies, PT targeted at smoking and substance abuse related behaviors appear to be the least successful with respect to the effectiveness of PT at promoting the desirable change in the domain. All three studies (100%) on smoking and substance abuse analyzed only reported partially positive results – a combination of positive with negative or no effect results. This is followed by disease management with 60% of all the studies reporting either negative or partially positive results. On the other hand, PT targeting dental health and sexual behaviors seem to be the most successful with respect to the effectiveness of PT at promoting the desirable change in the domain. All the studies related to dental health and sexual behaviors (100%) analyzed reported positive results. For eating related behaviora, 91% of all the studies reported fully positive results while for physical activity, 75% of all the studies reported fully positive results. Similarly, 67% of all the studies related to general health reported fully positive results. Finally, for the “Others” category which consists of health behaviors that appeared less frequently in our study, 73% of all the studies reported fully positive results. Figure 6 presents the comparative effectiveness of PT by targeted health domain.

Health behavior domain and motivational strategies employed. Some strategies were more frequently applied in one health and wellness domain than others. For example, tracking and monitoring is a common motivational and persuasive strategy in the physical activity and eating domain. As shown in Table 5, column 6, tracking and monitoring was employed 19 times and 7 times by PT targeting physical activity and eating behaviors respectively. Similarly, simulation and rehearsal is a popular strategy in the sexual behavior domain.

On a general note, tracking and monitoring; social support, sharing, and comparison; competition, leaderboard, and ranking; and rewards, points, and credits (listed in decreasing order of frequency) emerged as the top four strategies that are commonly employed by PT interventions in the physical activity domain. Similarly, for eating behavior tracking and monitoring; audio, visual, and textual feedback; and social support, sharing, and comparison (listed in decreasing order of frequency) emerged as the top three strategies. Finally, for dental behavior, audio, visual, and textual feedback followed by tracking and monitoring were the commonly employed strategies.

The variations in the popularity of the strategies across various health and wellness intervention domain may be due to the fact that some strategies are easier to operationalize in one domain than the other. However, there is no clear relationship between the strategies or the number of strategies employed in the PT design and persuasion outcome – PT effectiveness. This is probably due to many possible factors that could mediate the effectiveness of the strategies employed in PT including differences among the target population and the need to tailor the strategies to be appropriate for the target audience. The possible variations in operationalization and instantiations of the strategies is another possible factor.

Health behavior domain and behavior theories employed. Although most analyzed studies are not based on any known behavior theories or did not specify the theories that informed the design, for those that did, some theories seem to be more popular in one health and wellness domain than the other. As can be seen from from Table 5, column 7, theories such as the Transtheoretical model is more prevalent in the smoking and substance abuse, physical activity, and health eating domain. Similarly, goal setting theory is used mostly in physical activity and disease management domain. The variations in the popularity of the theories across various health and wellness intervention domains may be due to the fact that some theories are more suitable for some domain and can be easily operationalized than others. For instance, theories such as Transtheoretical model were developed in the context of smoking cessation. Although, it has since been proven effective for developing interventions targeting other health and wellness domains, it is more popular in the smoking cessation interventions as shown by this study. However, there is no clear relationship between the theory or the number of theories used to inform the PT and the persuasion outcome – PT effectiveness.

Targeted behavioral/psychological outcomes

With respect to the behavioral or psychological outcome targeted, as shown in Table 6, the studies are targeted at change in 12 distinct outcomes. Nearly half of all the studies (44%) are targeted at actual *Behavior* change (either promoting desirable behaviors or motivating change of undesirable behaviors). Seventeen percent (17%) were targeted at *Attitude* change and 17% at increasing *Motivation*, while an additional 7% were categorized as "*Other"* a category housing all studies that either did not specify the targeted behavioral outcome or included a targeted outcome unrelated to behavior. Most of the studies targeted more than one behavioral outcome and hence assessed the effectiveness of their PT on more than one behavioral outcome. As a result, some studies belong to more than one category in Table 6. The behavioral outcomes were not often measured using standardized instruments; creating or reinventing measurement instruments is a common trend among reviewed studies. Again, some studies that were targeted at actual behavior ended up evaluating the

Table 6. Behavioral/psychological outcomes targeted by persuasive technology.

Behavioral/Psychological Outcomes	Study	% of 85
Behavior	[4], [6], [7], [11], [12], [15], [17], [21]–[25], [27], [28], [32], [33], [35], [38], [41], [46], [48], [56], [58], [60], [62], [63], [65]–[67], [70]–[76], [79], [81], [83], [86], [88]	44%
Attitude	[5], [9], [10], [13], [17], [26], [30], [59], [65], [68], [76], [78], [84]	17%
Motivation	[6], [31], [37]–[40], [54], [57], [61], [64], [78]	17%
Awareness	[16], [43], [47], [50]–[52], [77], [84]	15%
Self-efficacy	[13], [17], [29], [30], [64], [69], [77], [84]	13%
Adherence and Compliance	[36], [34], [42], [49], [53], [55]	10%
Habit	[7], [8], [48]	5%
Knowledge	[8], [14], [80]	5%
Intention	[17], [69]	4%
Engagement and Acceptance	[32], [45]	4%
Belief and Perception	[10], [44], [89]	4%
Others	[10], [31], [82], [85]	7%

effectiveness of their systems by measuring some mediating psychological outcomes such as Attitude and Motivation because of the long evaluation period needed to establish actual change in behavior.

Study methodology used by persuasive technology

Data collection and analysis trend. Table 7 summarizes the methodology employed by the reviewed paper in evaluating their PT for health and wellness. Mixed method emerged as the dominant method. Of all the studies, 46% employed mixed method combining both quantitative and qualitative approaches in their study. This is followed by the quantitative approach, accounting for 39%. The most commonly used approach for collecting quantitative data was questionnaire/survey. A few studies (mostly physical activity motivating PT and games) collected additional quantitative data via logged data of user's behavior and system usage.[6,34,45,55–57,62,77,88–90] A fully qualitative approach is the least popular with only 15% of all the studies using the approach. The most frequently used qualitative methods are interviewing, focus-group discussion and observation of participants' behaviors and PT use.

Regarding the data analysis methods used in the reviewed studies, frequencies, percentages, and means and standard deviations are the most popular methods. ANOVA, Regression analysis and t-tests are the commonly used as inferential techniques. Content analysis was the most used qualitative data analysis methods.

Duration of evaluation. With respect to how long the PT were evaluated, the duration of evaluation varied substantially, ranging from 15 minutes to 3 years. However, some studies did not report how long their PT evaluation lasted. Only a few studies conducted a longitudinal evaluation of their PT[12,72,84] and the majority of the studies did not conduct a follow up study beyond the initial (feasibility) study. Therefore, it is difficult to establish long-term effects of PT for health and wellness from existing studies.

Table 7. Study methodologies used by persuasive technology.

Method	Studies with positive outcome	Studies with partially positive outcome	Studies with negative/other outcome	% of all
Quantitative	[5], [12], [13], [15], [21], [22], [24], [28], [29], [35], [41], [42], [44], [45], [48], [49], [51], [52], [54], [59], [64], [65], [67], [69], [70], [89]	[10], [74], [78], [80], [82]	[87]	39%
Qualitative	[7], [14], [33], [38]–[40], [43], [44], [46], [47]	[76]	[85], [86]	15%
Mixed Method	[4], [6], [8], [9], [16], [17], [23], [25]–[27], [30]–[32], [36], [37], [40], [43], [55], [57], [58], [60], [62], [63], [66], [68]	[10], [24], [28], [31], [71], [75], [76]	[15], [83], [84], [88]	46%

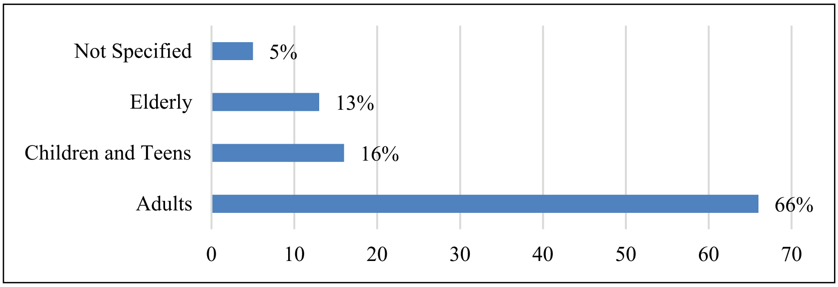


Figure 7. Targeted age demographic.

Study participants and sample size. Similar to the study duration, the sample size (number of participants in the evaluation of the PT) also varies greatly. The sample size ranges from 1 to 16,340 participants, with a mean sample size of 258 participants. However, one study did not report the total number of participants in their evaluation[61] and some studies are conducted in stages with the sample size and composition varying at each stage. In such cases, we report a combined sample size from all stages. Most of the study participants are recruited using a convenience sampling method and were recruited from academic communities or via online forums.

As can be seen from Figure 7, 66% of all the studies are targeted at adults, 16% involved children and teens, only 13% are targeted specifically at elderly people, and 5% of all the studies did not specify the target audience.

Discussion

Effectiveness of persuasive technology for health and wellness

Following from the reviewed literature, it can be concluded that persuasive technologies are effective at promoting various health and wellness related behavior with 92% of all the reviewed studies reporting some positive outcome (fully and partially positive) from PT use.

Although the majority of the studies (72%) are targeted at behavior and/or attitude change in line with the original conceptualization of PT by Fogg[91] (technology intended to change attitude and/or behavior), generally, the studies targeted and measured various other behavior-related or psychological outcomes beyond the conventional outcome of behavior and attitude. Similarly, some of the technologies are aimed at reinforcing and strengthening existing behavior (e.g., increase daily step count while others aimed at changing behavior (quitting smoking). This shows that persuasive technology has evolved over the years to encompass various practices that were not in the initial conceptualization.

The relationship between target health behavior domain (persuasion context) and persuasive technology outcome

There seem to be some variations in the effectiveness of the PT across various health and wellness domain. PT targeted at smoking and substance abuse related behaviors appear to be the least successful with respect to their overall effectiveness at promoting the desirable change in the domain. This is followed by PT targeting disease management. On the other hand, PT targeted at dental health; sexual behaviors; eating related behaviors; and physical activity (listed in decreasing order) seem to be the most successful with respect to the effectiveness of PT at promoting the desirable change in the domains. It is important to note that the effectiveness of the PT in various domains could be influenced by many factor including the operationalization of the persuasion strategies and behavior theories, the length and depth of evaluation, and the appropriateness of the PT for the target audience. Therefore, due to this methodological plurality and the heterogeneity of sample sizes and data, we are not able to draw strong conclusions about which persuasion contexts provided the most positive effects.

The relationship between behavior theory and persuasive technology outcome

Although, it is difficult to fully establish that using behavior theory to inform the design of PT influences their effectiveness due to the limited number of study in this review (85), the review results suggest that there may be some relationships. All the studies that are based on known theories but one[15] reported either fully positive or partially positive results. The study,[15] which failed employed an instantiation of negative reinforcement hence confirming that the use of any form of punishment as motivational strategy may backfire. Theories such as TTM and goal setting theory are more prevalent in the literature than others. However, the popularity of the theories vary across the domains. Theories such as TTM is common in the smoking and substance abuse, physical activity, and eating related domains.

The relationship between motivational strategies and persuasive technology outcome

Although several research efforts have been directed toward developing taxonomies for naming and classifying persuasive and motivational strategies,[91,92] there still exist many inconsistencies in both naming and operationalizing strategies in persuasive systems. In some cases, the only way that one could possibly identify the actual motivational strategies employed in the PT is by studying how they work. Again, while some strategies such as monitoring/tracking, feedback, and social support seem to be more frequently used than others, it does not seem that there is a relationship between the strategy employed and the success of the PT. This is probably because of the variations in the instantiation and framing of the strategies. For instance, feedback is often instantiated in different forms, including audio, visual, or text based feedback, in various degrees of granularity, and in one of two valences – positive or negative. This variation in instantiations is likely to impact effectiveness. The choice of how to instantiate and operationalize the strategies are solely based on PT designer's own discretion. Again, some strategies are more dominant in some health domains than the other. For example, monitoring and tracking strategy is a predominant strategy in physical activity and eating behavior motivating PT than other health wellness domains. Similarly, simulation and rehearse is a popular strategy in the sexual behavior domain.

General limitations and recommendations for future research

Based on the results of this review, we identified specific gaps in the literature and we offer suggestions for improvement and moving the field forward:

1. The PT literature lacks standardized approaches or tools for evaluating the effectiveness of PT and most existing evaluation approaches are based on subjective data which can be biased. Research into alternative assessment and evaluation techniques would enrich the PT community. In particular, the persuasive community would benefit greatly from research into objective evaluation approaches.
2. PT studies are limited in terms of effective integration of behavior theories and practice in their design. This is probably because most PT designers often lack the skills needed to translate theoretical determinants of behavior into technology design artefacts. We recommend that future research be focused on developing a comprehensive framework for translating theoretical determinants into technology design components.
3. Most PT employed more than one strategy in their design (see the appendix), therefore making it difficult to establish whether there is a relationship between the strategies and success of the PT. Research aimed at establishing the interactions between individual strategies and the success of PT either using sophisticated statistical techniques or by examining the effectiveness of the strategies in isolation would be vital to the community and would contribute in advancing the field.
4. Only a few studies have conducted longitudinal evaluations of the effectiveness of their PT.[12,72,84] We stress the need for research in this direction to establish the long-term effect of PT on health and wellness.
5. Only a few PT involved the target audience in their design.[30,79] We recommend that PT designers adopt the participatory design approach to enable the involvement of the target group(s) in deciding on the theories, strategies, and particular instantiation that will be suitable for the target audience and behavior.
6. Finally, there is a need for more PT to target diverse demographies such as older adults and children.

Conclusion

This paper provides a review of the effectiveness and trends of Persuasive Technology (PT) for health and wellness. The review results show that PT is a promising approach for promoting desirable behavior on a broad and heterogeneous range of health and wellness. However, lack of large-scale and longitudinal evaluations makes it impossible to establish the long-term impact of PT at promoting desirable behavior in the area of health and wellness. The review also highlighted PT trends and limitations of existing studies and suggested some improvements and future research direction.

Declaration of conflicting interests

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding

The author(s) received no financial support for the research, authorship, and/or publication of this article.

References

1. Orji R, Mandryk RL, Vassileva J, et al. Tailoring persuasive health games to gamer type. In: *Proceedings of the SIGCHI conference on human factors in computing systems (CHI '13)*, Paris, 27 April–2 May 2013, pp. 2467–2476. New York: ACM Press.
2. Orji R and Mandryk RL. Developing culturally relevant design guidelines for encouraging healthy eating behavior. *Int J Hum: Comput St* 2014; 72: 207–223.
3. Orji R, Vassileva J and Mandryk RL. Modeling the efficacy of persuasive strategies for different gamer types in serious games for health. *User Model User: Adap* 2014; 24: 453–498.
4. Lin JJ, Mamykina L, Lindtner S, et al. Fish 'n' steps: encouraging physical activity with an interactive computer game. In: Dourish P and Friday A (eds) *UbiComp 2006: ubiquitous computing*. Berlin: Springer, 2006, pp. 261–278.
5. Clinkenbeard D, Clinkenbeard J, Faddoul G, et al. What's your 2%? A pilot study for encouraging physical activity using persuasive video and social media. In: *Proceedings of the 9th international conference on persuasive technology*, Padua, 21–23 May 2014, pp. 43–55. Berlin: Springer.
6. Consolvo S, McDonald D, Toscos T, et al. Activity sensing in the wild: a field trial of ubifit garden. In: *Proceedings of the 26th annual SIGCHI conference on human factors in computing systems*, Florence, 5–10 April 2008, pp. 1797–1806. New York: ACM.
7. Pollak J, Gay G, Byrne S, et al. It's time to eat! Using mobile games to promote healthy eating. *IEEE Pervas Comput* 2010; 9: 21–27.
8. Grimes A, Kantroo V and Grinter RE. Let's play!: Mobile health games for adults. In: *Proceedings of the 12th ACM international conference on ubiquitous computing*, Copenhagen, 26–29 September 2010, pp. 241–250. New York: ACM.
9. Orji R, Vassileva J and Mandryk RL. LunchTime: a slow-casual game for long-term dietary behavior change. *Pers Ubiquit Comput* 2013; 17: 1211–1221.
10. Khaled R, Barr P, Biddle R, et al. Game design strategies for collectivist persuasion. In: *Proceedings of the 2009 ACM SIGGRAPH symposium on video games (Sandbox '09)*, New Orleans, LA, 3–7 August 2009, p. 31. New York: ACM.
11. Graham C, Benda P, Howard S, et al. heh-keeps me off the smokes...: probing technology support for personal change. In: *Proceedings of the 18th Australia conference on Computer-Human Interaction: Design: Activities, Artefacts and Environments,* 20 November 2006, pp. 221–228. ACM digital library.
12. Rosser BS, Oakes JM, Konstan J, et al. Reducing HIV risk behavior of MSM through persuasive computing: results of the men's INTernet Study (MINTS-II). *AIDS* 2010; 24: 2099–2107.
13. Roberto AJ, Zimmerman RS, Carlyle KE, et al. The effects of a computer-based pregnancy, STD, and HIV prevention intervention: a nine-school trial. *Health Commun* 2007; 21: 115–124.
14. Chang Y-C, Lo J-L, Huang C-J, et al. Playful toothbrush: ubicomp technology for teaching tooth brushing to kindergarten children. In: *Proceedings of the ACM conference on human factors in computing systems (CHI 2008)*, Florence, 5–10 April 2008, vol. 1, pp. 363–372. New York: ACM.
15. Nakajima T and Lehdonvirta V. Designing motivation using persuasive ambient mirrors. *Pers Ubiquit Comput* 2013; 17: 107–126.
16. Soler C, Zacarías A and Lucero A. Molarcropolis: a mobile persuasive game to raise oral health and dental hygiene awareness. In: *Proceedings of the international conference on advances in computer entertainment technology*, Athens, 29–31 October 2009, pp. 388–391. New York: ACM.

17. Orji R. *Design for behaviour change: a model-driven approach for tailoring persuasive technologies*. PhD Thesis, University of Saskatchewan, Saskatoon, SK, Canada, 2014.
18. Riff D, Lacy S and Fico F. *Analyzing media messages: using quantitative content analysis in research*. London, England: Routledge, 2014.
19. Elsevier BV, Scopus, http://www.scopus.com
20. Hamari J, Koivisto J and Pakkanen T. Do persuasive technologies persuade? A review of empirical studies. In: Spagnolli A, Chittaro L and Gamberini L (eds) *Persuasive technology*, vol. 8462. Berlin: Springer, 2014, pp. 118–136.
21. Khalil A and Abdallah S. Harnessing social dynamics through persuasive technology to promote healthier lifestyle. *Comput Hum Behav* 2013; 29: 2674–2681.
22. Young M. Twitter me: using micro-blogging to motivate teenagers to exercise. In: Winter R, Leon Zhao J and Aier S (eds) *Global perspectives on design science research*. Berlin: Springer, 2010, vol. 6105, pp. 439–448.
23. Van Leer E and Connor NP. Use of portable digital media players increases patient motivation and practice in voice therapy. *J Voice* 2012; 26: 447–453.
24. Salam SNA, Yahaya WAJW and Ali AM. Using persuasive design principles in motivational feeling towards children dental anxiety (CDA). In: Ploug T, Hasle P and Oinas-Kukkonen H (eds) *Persuasive technology* (Lecture notes in computer science (including sub series lecture notes in artificial intelligence lecture notes bioinformatics). Berlin: Springer, 2010, pp. 223–237.
25. Mutsuddi AU and Connelly K. Text messages for encouraging physical activity: are they effective after the novelty effect wears off? In: *Proceedings of the 2012 6th international conference on pervasive computing technologies for healthcare (Pervasive Health) and workshops*, San Diego, CA, 21–24 May 2012, pp. 33–40. New York: IEEE.
26. Looije R, Cnossen F and Neerincx MA. Incorporating guidelines for health assistance into a socially intelligent robot. In: *Proceedings of the 15th IEEE international symposium on robot and human interactive communication*, Hatfield, 6–8 September 2006, pp. 515–520. New York: IEEE.
27. Lim BY, Shick A, Harrison C, et al. Pediluma: motivating physical activity through contextual information and social influence. In: *Proceedings of the 5th international conference on tangible, embedded, and embodied interaction (TEI 2011)*, Funchal, 23–26 January 2011, pp. 173–180. New York: ACM.
28. Lee MK, Kiesler S and Forlizzi J. Mining behavioral economics to design persuasive technology for healthy choices. In: *Proceedings of the SIGCHI conference on human factors in computing systems*, Vancouver, BC, Canada, 7–12 May 2011, pp. 325–334. New York: ACM.
29. Lacroix J, Saini P and Goris A. Understanding user cognitions to guide the tailoring of persuasive technology-based physical activity interventions. In: *Proceedings of the 4th international conference on persuasive technology*, Claremont, CA, 26–29 April 2009, pp. 1–8. New York: ACM.
30. Kroes L and Shahid S. Empowering young adolescents to choose the healthy lifestyle: a persuasive intervention using mobile phones. In: *Proceedings of the 15th international conference on human-computer interact (HCI)*, Las Vegas, NV, 21–26 July 2013, pp. 117–126. Berlin: Springer.
31. Kehr F, Hassenzahl M, Laschke M, et al. A transformational product to improve self-control strength: the chocolate machine. In: *Proceedings of the 2012 annual conference on human factors in computing systems (CHI '12)*, Austin, TX, 5–10 May 2012, pp. 689–694. New York: ACM.
32. Gasca E, Favela J and Tentori M. Persuasive virtual communities to promote a healthy lifestyle among patients with chronic diseases. In: Briggs RO, Antunes P, de Vreede G-J, et al. (eds) *Groupware: design, implementation, and use*. Berlin: Springer, 2008, pp. 74–82.
33. Fabri M, Wall A and Trevorrow P. Changing eating behaviors through a cooking-based website for the whole family (Healthy eating behaviors). In: Marcus A (ed.) *Design, user experience, and usability: user experience in novel technological environments*. Berlin: Springer, 2013, pp. 484–493.
34. De Oliveira R, Cherubini M and Oliver N. MoviPill: improving medication compliance for elders using a mobile persuasive social game. In: *Proceedings of the 12th ACM international conference on ubiquitous computing*, Copenhagen, 26–29 September 2010, pp. 251–260. New York: ACM.

35. Chiu M, Chang S, Chang Y, et al. Playful bottle: a mobile social persuasion system to motivate healthy water intake. In: *Proceedings of the 11th international conference on ubiquitous computing*, Orlando, FL, 30 September–3 October 2009, pp. 184–194. New York: ACM.
36. Bhatnagar N, Sinha A, Samdaria N, et al. Biometric monitoring as a persuasive technology: ensuring patients visit health centers in India's slums. In: Bang M and Ragnemalm EL (eds) *Persuasive technology: design for health and safety* (Lecture notes in computer science (including sub series lecture notes in artificial intelligence lecture notes bioinformatics)). Berlin: Springer, 2012, vol. 7284, pp. 169–180.
37. Berque D, Burgess J, Billingsley A, et al. Design and evaluation of persuasive technology to encourage healthier typing behaviors. In: *Proceedings of the 6th international conference on persuasive technology: persuasive technology and design: enhancing sustainability and health*, Columbus, OH, 2–5 June 2011, pp. 1–10. New York: ACM.
38. Berkovsky S, Freyne J and Coombe M. Physical activity motivating games: be active and get your own reward. *ACM T Comput: Hum Int* 2012; 19: 1–41.
39. Arteaga SM, Kudeki M, Woodworth A, et al. Mobile system to motivate teenagers' physical activity. In: *Proceedings of the 9th international conference on interaction design and children*, Barcelona, 9–12 June 2010, pp. 1–10. New York: ACM.
40. Eyck A, Geerlings K, Karimova D, et al. Effect of a virtual coach on athletes' motivation. In: IJsselsteijn WA, de Kort YAW, Midden C, et al. (eds) *Persuasive technology*, vol. 3962 (Lecture notes in computer science (including sub series lecture notes in artificial intelligence lecture notes bioinformatics)). Berlin: Springer, 2006, pp. 158–161.
41. Foster D, Linehan C and Lawson S. Motivating physical activity at work: using persuasive social media extensions for simple mobile devices. In: *Proceedings of the 14th international academic MindTrek conference: envisioning future media environments (MobileHCI)*, 6 October 2010, pp. 11–14. ACM digital library.
42. Bickmore T, Mauer D, Crespo F, et al. Persuasion, task interruption and health regimen adherence. In: De Kort Y, IJsselsteijn W, Midden C, et al. (eds) *Persuasive technology*. Berlin: Springer, 2007, pp. 1–11.
43. Obermair C, Reitberger W, Meschtscherjakov A, et al. PerFrames: persuasive picture frames for proper posture. In: Oinas-Kukkonen H, Hasle P, Harjumaa M, et al. (eds) *Persuasive technology*, vol. 5033(Lecture notes in computer science (including sub series lecture notes in artificial intelligence lecture notes bioinformatics)). Berlin: Springer, 2008, pp. 128–139.
44. Parmar V and Keyson D. Persuasive technology for shaping social beliefs of rural women in India: an approach based on the theory of planned behaviour. In: *Proceedings of the 3rd international conference on persuasive 2008*, Oulu, 4–6 June 2008, pp. 104–115. Berlin: Springer.
45. Gasser R, Brodbeck D, Degen M, et al. Persuasiveness of a mobile lifestyle coaching application using social facilitation. In: IJsselsteijn WA, de Kort YAW, Midden C, et al. (eds) *Persuasive technology*. Berlin: Springer, 2006, pp. 27–38.
46. Fritz T, Huang EM, Murphy GC, et al. Persuasive technology in the real world. In: *Proceedings of the 32nd annual ACM conference on human factors in computing systems (CHI '14)*, Toronto, ON, Canada, 26 April–1 May 2014, pp. 487–496. New York: ACM.
47. Sohn M and Lee J. UP health: ubiquitously persuasive health promotion with an instant messaging system. In : *Proceedings of the extended abstracts on human factors in computing systems (CHI '07)*, San Jose, CA, 28 April–3 May 2007, pp. 2663–2668. New York: ACM.
48. Toscos T, Faber A, An S, et al. Chick clique: persuasive technology to motivate teenage girls to exercise. In: *Proceedings of the extended abstracts on human factors in computing systems (CHI '06)*, Montréal, QC, Canada, 22–27 April 2006, pp. 1873–1878. New York: ACM.
49. Sterns AA and Mayhorn CB. Persuasive pillboxes: improving medication adherence with personal digital assistants. In: IJsselsteijn WA, de Kort YAW, Midden C, et al. (eds) *Persuasive technology*. Berlin, Heidelberg: Springer, 2006, pp. 195–198.
50. Chi P-YP, Chu H-H, Chen J, et al. Enabling nutrition-aware cooking in a smart kitchen. In: *Proceedings of the extended abstracts on human factors in computing systems (CHI '07)*, San Jose, CA, 28 April–3 May 2007, pp. 2333–2338. New York: ACM.

51. Golsteijn C, Van Den Hoven E, Geurts S, et al. BLB: a persuasive and interactive installation designed to improve well-being. In: Oinas-Kukkonen H, Hasle P, Harjumaa M, et al. (eds) *Persuasive technology*, vol. 5033 (Lecture notes in computer science (including sub series lecture notes in artificial intelligence lecture notes bioinformatics)). Berlin: Springer, 2008, pp. 262–265.
52. Fujinami K and Riekki J. A case study on an ambient display as a persuasive medium for exercise awareness. In: *Proceedings of the 3rd international conference on persuasive technology*, Oulu, 4–6 June 2008, vol. 5033, pp. 266–269. Berlin: Springer.
53. Munson SA, Lauterbach D, Newman MW, et al. Happier together: integrating a wellness application into a social network site. In: Ploug T, Hasle P and Oinas-Kukkonen H (eds) *Persuasive technology*, vol. 6137 (Lecture notes in computer science (including sub series lecture notes in artificial intelligence lecture notes bioinformatics)). Berlin: Springer, 2010, pp. 27–39.
54. Spruijt-Metz D, Nguyen-Michel ST, Goran MI, et al. Reducing sedentary behavior in minority girls via a theory-based, tailored classroom media intervention. *Int J Pediatr Obes* 2008; 3: 240–248.
55. Tsai CC, Lee G, Raab F, et al. Usability and feasibility of PmEB: a mobile phone application for monitoring real time caloric balance. *Mobile Netw Appl* 2007; 12: 173–184.
56. Fujiki Y, Kazakos K, Puri C, et al. NEAT-o-games: blending physical activity and fun in the daily routine. *Comput Entertain* 2008; 6: 1–22.
57. Kaipainen K, Mattila E, Kinnunen ML, et al. Facilitation of goal-setting and follow-up in an internet intervention for health and wellness. In: Ploug T, Hasle P and Oinas-Kukkonen H (eds) *Persuasive technology*, vol. 6137 (Lecture notes in computer science (including sub series lecture notes in artificial intelligence lecture notes bioinformatics)). Berlin: Springer, 2010, pp. 238–249.
58. Chatterjee S, Byun J, Pottathil A, et al. Persuasive sensing: a novel in-home monitoring technology to assist elderly adult diabetic patients. In: Bang M and Ragnemalm EL (eds) *Persuasive technology: design for health and safety*, vol. 7284 (Lecture notes in computer science). Berlin: Springer, 2012, pp. 31–42.
59. Chittaro L and Sioni R. Turning the classic snake mobile game into a location–based exergame that encourages walking. In: Bang M and Ragnemalm EL (eds) *Persuasive technology: design for health and safety*, vol. 7284 (Lecture notes in computer science (including sub series lecture notes in artificial intelligence lecture notes bioinformatics)). Berlin: Springer, 2012, pp. 43–54.
60. Langrial S and Oinas-Kukkonen H. Less fizzy drinks: a multi-method study of persuasive reminders. In: Bang M and Ragnemalm EL (eds) *Persuasive technology: design for health and safety: 7th international conference on persuasive technology 2012, Linköping, 6–8 June 2012*, vol. 7284. Berlin: Springer, 2012, pp. 256–261.
61. Peeters M, Megens C, Van Den Hoven E, et al. Social stairs: taking the piano staircase towards long-term behavioral change. In: Berkovsky S and Freyne J (eds) *Persuasive technology*, vol. 7822 (Lecture notes in computer science (including sub series lecture notes in artificial intelligence lecture notes bioinformatics)). Berlin: Springer, 2013, pp. 174–179.
62. Chen YX, Chiang SS, Chih SY, et al. Opportunities for persuasive technology to motivate heavy computer users for stretching exercise. In: Spagnolli A, Chittaro L and Gamberini L (eds) *Persuasive technology*, vol. 8462. Berlin: Springer, 2014, pp. 25–30.
63. Langrial S, Oinas-Kukkonen H, Lappalainen P, et al. Managing depression through a behavior change support system without face-to-face therapy. In: Spagnolli A, Chittaro L and Gamberini L (eds) *Persuasive technology*, vol. 8462. Berlin: Springer, 2014, pp. 155–166.
64. Kopf LM, Graetzer S and Huh J. Videos influence behavior change measures for voice and speech in individuals with Parkinson's disease. *Proc Wirel Health*. Epub ahead of print October 2015. DOI: 10.1145/2811780.2811932.
65. Takeuchi T, Fujii T, Narumi T, et al. Considering individual taste in social feedback to improve eating habits. In: *Proceedings of the 2015 IEEE international conference on multimedia and expo workshops (ICMEW)*, Turin, 29 June–3 July 2015, pp. 1–6. New York: IEEE.
66. Cornejo R, Hernandez D, Tentori M, et al. Casual gaming to encourage active ageing. *Lat Am Trans IEEE* 2015; 13: 1940–1950.
67. Lo J, Lin T, Chu H, et al. Playful tray: adopting ubicomp and persuasive techniques into play-based occupational therapy for reducing poor eating behavior in young children. In: *Proceedings of the inter-*

national conference on ubiquitous computing (UbiComp '07), Innsbruck, 16–19 September 2007, pp. 38–55. Berlin: Springer.

68. Gerber BS, Stolley MR, Thompson AL, et al. Mobile phone text messaging to promote healthy behaviors and weight loss maintenance: a feasibility study. *Health Informatics J* 2009; 15: 17–25.
69. Peng W. Design and evaluation of a computer game to promote a healthy diet for young adults. *Health Commun* 2009; 24: 115–127.
70. Kaptein M, De Ruyter B, Markopoulos P, et al. Adaptive persuasive systems. *ACM Trans Interact Intell Syst* 2012; 2: 1–25.
71. Kim S, Kientz JA, Patel SN, et al. Are you sleeping? Sharing portrayed sleeping status within a social network. In: *Proceedings of the 2008 ACM conference on computer supported cooperative work*, San Diego, CA, 8–12 November 2008, pp. 619–628. New York: ACM.
72. VanDeMark NR, Burrell NR, Lamendola WF, et al. An exploratory study of engagement in a technology-supported substance abuse intervention. *Subst Abuse Treat Prev Policy* 2010; 5: 10.
73. Toscos T, Faber A, Connelly K, et al. Encouraging physical activity in teens can technology help reduce barriers to physical activity in adolescent girls? In: *Proceedings of the 2008 2nd international conference on pervasive computing technologies for healthcare*, Tampere, 30 January–1 February 2008, vol. 3, pp. 4–7. New York: IEEE.
74. Sakai R, Van Peteghem S, van de Sande L, et al. Personalized persuasion in ambient intelligence: The apstairs system. In: *Ambient Intelligence* 16 Novemebr 2011, pp. 205–209. Berlin Heidelberg: Springer.
75. Munson S and Consolvo S. Exploring goal-setting, rewards, self-monitoring, and sharing to motivate physical activity. In: *Proceedings of the 2012 6th international conference on pervasive computing technologies for healthcare and workshops*, San Diego, CA, 21–24 May 2012, pp. 25–32. New York: IEEE.
76. Mintz J, Branch C, March C, et al. Key factors mediating the use of a mobile technology tool designed to develop social and life skills in children with Autistic Spectrum Disorders. *Comput Educ* 2012; 58: 53–62.
77. Kaplan B, Farzanfar R and Friedman RH. Personal relationships with an intelligent interactive telephone health behavior advisor system: a multimethod study using surveys and ethnographic interviews. *Int J Med Inform* 2003; 71: 33–41.
78. Jeen Y, Han J, Kim H, et al. Persuasive interaction strategy for self diet system: exploring the relation of user attitude and intervention by computerized systematic methods. In: *Proceedings of the 12th international conference on human-computer interaction, HCI international 2007*, Beijing, China, 22–27 July 2007, pp. 450–458. Berlin: Springer.
79. Albaina IM, Visser T, Mast CA, et al. Flowie: a persuasive virtual coach to motivate elderly individuals to walk. In: *Proceedings of the 2009 3rd international conference on pervasive computing technologies for healthcare*, London, 1–3 April 2009, pp. 1–7. New York: IEEE.
80. Looije R, Neerincx MA and Cnossen F. Persuasive robotic assistant for health self-management of older adults: design and evaluation of social behaviors. *Int J Hum: Comput St* 2010; 68: 386–397.
81. Knipscheer K, Nieuwesteeg J and Oste J. Persuasive story table: promoting exchange of life history stories among elderly in institutions. In: IJsselsteijn WA, De Kort Y, Midden C, et al. (eds) *Persuasive technology*, vol. 3962 (Lecture notes in computer science (including sub series lecture notes in artificial intelligence lecture notes bioinformatics)). Berlin: Springer, 2006, pp. 191–194.
82. McCalley T and Mertens A. The pet plant: developing an inanimate emotionally interactive tool for the elderly. In: De Kort Y, IJsselsteijn W, Midden C, et al. (eds) *Persuasive technology*. Berlin: Springer, 2007, pp. 68–79.
83. Kientz JA, Choe EK, Birch B, et al. Heuristic evaluation of persuasive health technologies. In: *Proceedings of the ACM international conference health informatics (IHI '10)*, Washington, DC, 11–12 November 2010, p. 555. New York: ACM.
84. Goessens BMB, Visseren FLJ, Geerts AC, et al. Self-management of vascular patients activated by the Internet and nurses: rationale and design. In: IJsselsteijn WA, de Kort YAW, Midden C, et al. (eds) *Persuasive technology*, vol. 3962. Berlin: Springer, 2006, pp. 162–166.
85. McCreadie C, Raper J, Gunesh A, et al. Persuasive technology for leisure and health: development of a personal navigation tool. In: IJsselsteijn WA, de Kort YAW, Midden C, et al. (eds) *Persuasive technology*, vol. 3962. Berlin: Springer, 2006, pp. 187–190.

86. Harjumaa M. Understanding persuasive software functionality in practice: a field trial of polar FT60. In: *Proceedings of the 4th international conference on persuasive technology*, Claremont, CA, 26–29 April 2009. New York: ACM.
87. Burigat S and Chittaro L. Designing a mobile persuasive application to encourage reduction of users' exposure to cell phone RF emissions. In: Spagnolli A, Chittaro L and Gamberini L (eds) *Persuasive technology*, vol. 8462. Berlin: Springer, 2014, pp. 13–24.
88. Zwinderman MJ, Shirzad A, Ma X, et al. Phone row: a smartphone game designed to persuade people to engage in moderate-intensity physical activity. In: Bang M and Ragnemalm EL (eds) *Persuasive technology: design for health and safety* (Lecture notes in computer science). Berlin: Springer, 2012, pp. 55–66.
89. Adams AT, Costa J, Jung MF, et al. Mindless computing: designing technologies to subtly influence behavior. In: *Proceedings of the 2015 ACM international joint conference on pervasive and ubiquitous computing (UbiComp 2015)*, Osaka, Japan, 7–11 September 2015, pp. 719–730. New York: ACM.
90. Consolvo S, Markle K, Patrick K, et al. Designing for persuasion: mobile services for health behavior change. In: *Proceedings of the 4th international conference on persuasive technology*, Claremont, CA, 26–29 April 2009, p. 1. New York: ACM.
91. Fogg BJ. *Persuasive technology: using computers to change what we think and do*. San Francisco, CA: Morgan, 2009.
92. Torning K and Oinas-Kukkonen H. Persuasive system design: state of the art and future directions. In: *Proceedings of the 4th international conference on persuasive technology*, Claremont, CA, 26–29 April 2009. New York: ACM.

Appendix 1. A comprehensive overview of persuasive technology for health and wellness.

Authors	System/Project Name	Health Domain	Technology	Motivational Strategy	Behavior Theory	Country	Reference
Kim et al. (2008)	BuddyClock	Sleeping	Mobile	Feedback, status sharing	None	USA	[71]
Graham et al. (2006)	QuitCoach	Smoking	Web	Social support, Feedback and Advice	Transtheoretical model	Australia	[11]
Zwinderman et al. (2012)	Phone Row	Physical activity	Mobile game	Competition and leaderboard, feedback, and social comparison via Facebook	None	Netherlands	[88]
Young (2010)	Twitter Me	Physical activity	Combination of web, phone, and pedometer	Tracking and monitoring, sharing, Competition, praise, goal setting, reward, social comparison, persuasive text messaging	None	Netherlands	[22]
VanDeMark et al. (2010)	E-TREAT	Substance abuse	Computer-based application	Tailored persuasive messages, Individualized feedback, personalized coaching and support, suggestion	Transtheoretical model	USA	[72]
Van Leer and Connor (2012)	Mobile video	Voice therapy	Mobile	Video-based persuasion	Goal setting theory, social cognitive theory	USA	[23]
Toscos et al. (2008)	Mobile app.	Physical activity	Mobile	Social support, Persuasive text message, tracking and monitoring, sharing achievements	None	USA	[73]
Salam et al. (2010)	PMLE	Dental anxiety	CD ROM	Virtual Rehearsal, similarity, praise, and social learning	None	Malaysia	[24]
Sakai et al. (2011)	APStair	Physical activity	Publicly displayed screen	Authority, commitment, and consensus	None	Netherlands	[74]
Nakajima and Lehdonvirta (2013)	Persuasive Art	Physical activity	Ambient mirror	Tracking and monitoring, Reward, persuasive visual art, feedback, empathetic expressions	Goal setting theory, reinforcement theory	Japan	[15]
Nakajima and Lehdonvirta (2013)	Virtual Aquarium	Dental hygiene - Tooth brushing	Ambient mirror	Goal and objective, tracking and monitoring, visual feedback, positive reinforcement	Reinforcement theory, information theory	Japan	[15]
Mutsuddi and Connelly (2012)	Text messaging app	Physical activity	Desktop text messaging app	Goal and objective, reward, sharing testimonies	Transtheoretical model	USA	[25]
Munson and Consolvo (2012)	GoalPost and GoalLine	Physical activity	Mobile	Goal and objective, self-monitoring, sharing, reminder	Goal setting theory	USA	[75]

(Continued)

Appendix 1. (Continued)

Authors	System/Project Name	Health Domain	Technology	Motivational Strategy	Behavior Theory	Country	Reference
Mintz et al. (2012)	HANDS iProject	Autism	Mobile	Source credibility - expertise and trustworthiness, reward	None	Denmark, Sweden, Hungary, and UK	[76]
Looije et al. (2006)	iCat-personal assistant robot	Diabetes	Robot	Emoticons and visual expressions	Unified theory of acceptance and use of technology	Netherlands	[26]
Lim et al. (2011)	Pediluma	Physical activity	Physical activity tracker	Tracking and monitoring, visual feedback, reward, sharing	Transtheoretical model	USA	[27]
Lee et al. (2011)	Snackbot	Eating	Robot	default	None	USA	[28]
Lee et al. (2011)	Snack ordering site	Eating	Web	Default and Information	None	USA	[28]
Lacroix et al. (2009)	Activity monitor	Physical activity	Wearable device	Tracking and monitoring, visual feedback.	Self-determination	Netherlands	[29]
Kroes and Shahid (2013)	Powerfood	Eating	Mobile	Social Influence, feedback, comparison, reminder, points	Transtheoretical model, technology acceptance model	Netherlands	[30]
Kehr et al. (2012)	Chocolate Machine	Eating	Chocolate machine	Not specified	Ego depletion theory	Germany	[31]
Kaplan et al. (2003)	TLC-Computer-based telecommunication system	Chronic disease management	Computer-based application	Goal and objective, reminder	Goal setting theory	USA	[77]
Jeen et al. (2007)	Self-Diet System	Eating	Web	Source credibility, praise, negative reinforcement, social facilitation	Social learning theory	Korea	[78]
Foster et al. (2010)	Step Matron	Physical activity	Facebook app and pedometer	Competition, comparison	None	UK	[41]
Gasca et al. (2008)	pHealthNet	Eating and physical activity	Pedometer, mobile and web	Social support, collaboration, commitment	None	Mexico	[32]
Fabri et al. (2013)	Cooking Website	Eating	Web	Tracking and monitoring, Persuasive images	Transtheoretical model	UK	[33]
Consolvo et al. (2008)	UbiFit Garden	Physical activity	Mobile and activity sensor	Tracking and monitoring, visual feedback, reward, goal and objective progress display, positive reinforcement	Transtheoretical model	USA	[6]
Grimes et al. (2010)	OrderUp	Eating	Mobile game	Reward	Transtheoretical model	USA	[8]

Appendix 1. (Continued)

Authors	System/Project Name	Health Domain	Technology	Motivational Strategy	Behavior Theory	Country	Reference
De Oliveira et al. (2010)	MoviPill	Adherent to prescriptions	Mobile game	Tracking and monitoring, reward and point, reminder, competition	None	Spain	[34]
Chiu et al. (2009)	Playful bottle	Water intake (eating)	Mobile enabled Tracking and game	Tracking and monitoring, reminder, social support	Social conformity theory	Taiwan	[35]
Chang et al. (2008)	Playful toothbrush	Dental Hygiene - Tooth brushing	Game	Tracking and monitoring, visual and audio feedback, suggestion, praise	Teaching-learning theory	Taiwan	[14]
Bhatnagar et al. (2012)	Biometric system	Chronic disease management - tuberculosis	Biometric tracker	Tracking and monitoring	None	India	[36]
Berque et al. (2011)	Typing tracker	Healthy typing - repetitive strain injury	Computer-based application	Tracking and monitoring, reminder, visual and auditory feedback	None	USA	[37]
Berkovsky et al. (2012)	PLAY MATE!	Physical activity	Game	Tracking and monitoring, reward	Reinforcement theory, premack's principle	Australia	[38]
Arteaga et al. (2010)	Mobile app	Physical activity	Mobile game	Competition and Reward	Theory of planned behavior, theory of meaning of behavior, personality theory	USA	[39]
Pollak et al. (2010)	Time to Eat!	Eating	Mobile game	Reminder, Control, positive and negative feedback	None	USA	[7]
Soler et al. (2009)	Molarcropolis	Dental hygiene - tooth brushing	Mobile game	Visual feedback	None	Greece	[16]
Khaled et al. (2009)	Smoke?	Smoking	Computer game	Tracking and monitoring	None	New Zealand	[10]
Eyck et al. (2006)	Cycling system	Physical activity	Virtual coach	Tailoring and personalization, tunneling, praise	None	Netherlands	[40]
Albaina et al. (2009)	Flowie	Physical activity	Touch-screen, photo frame and pedometer	Tracking and monitoring, commitment and consistency, goal and objective	Goal setting, classic learning theory	Netherlands	[79]
Rosser et al. (2010)	MINTS-II	HIV/STI	Web	Customization, tunneling, simulation	Sexual health model	USA	[12]

(Continued)

Appendix 1. (Continued)

Authors	System/Project Name	Health Domain	Technology	Motivational Strategy	Behavior Theory	Country	Reference
Adams et al. (2015)	Mindless plate	Eating	Sensing plate	Tracking and monitoring, Competition and leaderboard, comparison,	None	USA	[89]
Kientz et al. (2010)	MyPyramid Blast Off	Eating and physical activity	Computer game	Praise, visual feedback	None	USA	[83]
Bickmore et al. (2007)	Wrist-rest agent	Wrist rest	PDA-based social agent	Audio alert	None	USA	[42]
Obermair et al. (2008)	perFrame	Healthy posture	Interactive picture frame	Feedback	None	Austria	[43]
Parmar and Keyson (2008)	PHI	Maternal health and menses	Computer-based application	Social cues	Theory of planned behavior	India	[44]
Gasser et al. (2006)	Mobile app	Eating and physical activity	Mobile and web	Social facilitation	None	Switzerland	[45]
Fritz et al. (2014)	Activity sensor	Physical activity	Activity sensing devices	Tracking and monitoring, reward, goal, conditioning, social-sharing	None	USA	[46]
Sohn and Lee (2007)	UP Desk	Physical activity, smoking	PD text messaging	Tracking and monitoring, competition and leaderboard, cooperation, goal and objective, reward	None	Korea	[47]
Looije et al. (2010)	Robotic assistant	Diabetes self-management	Computer-based application	Not specified	Unified theory of acceptance and use of technology, five factor model	Netherlands	[80]
Toscos et al. (2006)	Chick Clique	Physical activity	Mobile and pedometer	Tracking and monitoring, sharing, comparison, positive feedback	None	USA	[48]
Goessens et al. (2006)	SPAIN pilot-study	Patient's self-management	Computer-based application	Feedback, social support	None	Netherlands	[84]
McCreadie et al. (2006)	Personal Navigation tool	Physical activity	Mobile app	Tracking and monitoring, feedback	None	UK	[85]
Sterns and Mayhorn (2006)	PDA Pillbox	Medication adherence	PDA pillbox	Tracking and monitoring, reminder	None	USA	[49]
Knipscheer et al. (2006)	Persuasive story table	Loneliness and depression	Digital story telling table	Cooperation and collaboration	None	Netherlands	[81]
McCalley and Mertens (2007)	Pet plant	General health	Digital pet plant	Not specified	None	Netherlands	[82]

Appendix 1. (Continued)

Authors	System/Project Name	Health Domain	Technology	Motivational Strategy	Behavior Theory	Country	Reference
Chi et al. (2007)	Digital kitchen	Eating	Calorie aware kitchen	Tracking and monitoring, feedback	None	Taiwan	[50]
Golsteijn et al. (2008)	BLB	General well-being.	Persuasive bulbs	Not specified	None	Netherlands	[51]
Fujinami and Riekki (2008)	Ambient Mirror	Physical activity	Ambient mirror display	Competition, cooperation and collaboration	None	Japan	[52]
Harjumaa et al. (2009)	FT60	Physical activity	Heart rate monitor	Tracking and monitoring, personalization, reduction, praise, reward, reminder, credibility, goal and objective	None	Finland	[86]
Munson et al. (2010)	3GT	General well-being.	Social network	Reminder, sharing	None	USA	[53]
Roberto et al. (2007)	Computer-based application	Pregnancy, STD, and HIV prevention	Computer-based application	Not specified	Parallel process model	USA	[13]
Spruijt-Metz et al. (2008)	Get Moving!	Physical activity	Computer-based application	Not specified	Self-determination theory, theory of meanings of behavior	USA	[54]
Tsai et al. (2007)	PmEB mobile	Eating	Mobile app	Tracking and monitoring, reminder	None	USA	[55]
Fujiki et al. (2008)	NEAT-o-Games	Physical activity	Mobile game	Tracking and monitoring, competition	None	USA	[56]
Kaipainen et al. (2010)	GoodLife	Stress management	Web	Goal and objective, Tunneling	Goal setting theory, Transtheoretical model, cognitive behavior theory, acceptance and commitment therapy	Finland	[57]
Chatterjee et al. (2012)	In-home monitor	Diabetes self-management	Environmental and body-wearable sensors	Tracking and monitoring, reminder	None	USA	[58]
Chittaro and Sioni (2012)	LocoSnake game	Physical activity	Mobile game	Reward	None	Italy	[59]
Langrial and Oinas-Kukkonen (2012)	Web app	Eating	Web	Reminder, tracking	None	Finland	[60]

(Continued)

Appendix 1. (Continued)

Authors	System/Project Name	Health Domain	Technology	Motivational Strategy	Behavior Theory	Country	Reference
Peeters et al. (2013)	Social Stairs	Physical activity	Intelligent musical staircase	Reward	None	Netherlands	[61]
Burigat and Chittaro (2014)	Mobile app	Earphone use	Mobile app	Reminder, feedback, positive/negative reinforcement, progress	None	Italy	[87]
Chen et al. (2014)	SP-Stretch, Social Persuasion System	Physical activity	Mobile sensing and game	Tracking and monitoring, competition	None	Taiwan	[62]
Clinkenbeard et al. (2014)	Social network app	Physical activity	Social network	Tunneling, reduction, Suggestion	None	USA	[5]
Langrial et al. (2014)	Good Life Compass	Depression	Web	Reminder, rehearsal	None	Finland	[63]
Kopf et al. (2015)	Pakinson's Disease related videos	Parkinson's disease	Video	Not specified	Transtheoretical model	USA	[64]
Takeuchi et al. (2015)	Meal-sharing social media application	Eating	Social network	Feedbacks, sharing	None	Japan	[65]
Cornejo et al. (2015)	GuessMyCaption	Active ageing	Ambient casual game	Visual and audio feedbacks, sharing	None	Mexico	[66]
Orji et al. (2013)	LunchTime	Eating	Web	Goal and objective, reward, social influence, feedback	Transtheoretical model, knowledge-attitude-behavior theory	Canada	[9]
Lin et al. (2006)	Fish'n'Steps	Physical activity	Animated virtual fish	Goal and objective, tracking and monitoring, visual feedback, progress, sharing, competition, cooperation	Transtheoretical model	USA	[4]
Orji (2014)	JunkFood Alien	Eating	Phone-based game	Competition and Reward	None	Canada	[17]
Lo et al. (2007)	Playful Tray	Eating	Digital playful tray and games	Tracking and monitoring, visual feedback	None	Taiwan	[67]
Gerber et al. (2009)	Text messaging app	Eating and physical activity	Mobile	Personalization and tailoring	None	USA	[68]
Peng (2009)	RightWay Café	Eating	Computer game	Feedback, points, rehearsal	Theory of reasoned action, social cognitive theory, health belief model	USA	[69]
Kaptein et al. (2012)	Text messaging app	Eating - snacking	Mobile	Tracking and monitoring, liking, consensus, consistency, authority	None	Netherlands	[70]
Khalil and Abdallah (2013)	SET UP	Physical activity	Mobile	Tracking and monitoring, sharing	Theory of reasoned action	United Arab Emirate	[21]

Health Informatics Journal

Article

Health Informatics Journal
2018, Vol. 24(1) 92–102
© The Author(s) 2016
Reprints and permissions:
sagepub.co.uk/journalsPermissions.nav
DOI: 10.1177/1460458216653242
journals.sagepub.com/home/jhi
SAGE

Strategies to improve effectiveness of physical activity coaching systems: Development of personas for providing tailored feedback

Reinoud Achterkamp
Roessingh Research and Development, The Netherlands; University of Twente, The Netherlands

Marit GH Dekker-Van Weering
Roessingh Research and Development, The Netherlands

Richard MH Evering
Saxion University of Applied Sciences, The Netherlands

Monique Tabak, Josien G Timmerman, Hermie J Hermens and Miriam MR Vollenbroek-Hutten
Roessingh Research and Development, The Netherlands; University of Twente, The Netherlands

Abstract

Mobile physical activity interventions can be improved by incorporating behavioural change theories. Relations between self-efficacy, stage of change, and physical activity are investigated, enabling development of feedback strategies that can be used to improve their effectiveness. A total of 325 healthy control participants and 82 patients wore an activity monitor. Participants completed a self-efficacy or stage of change questionnaire. Results show that higher self-efficacy is related to higher activity levels. Patients are less active than healthy controls and show a larger drop in physical activity over the day. Patients in the maintenance stage of change are more active than patients in lower stages of change, but show an equally large drop in level of physical activity. Findings suggest that coaching should at least be tailored to level of self-efficacy, stage of change, and physical activity pattern. Tailored coaching strategies are developed, which suggest that increasing self-efficacy of users is most important. Guidelines are provided.

Keywords

coaching, feedback, physical activity, self-efficacy, stage of change

Corresponding author:
Reinoud Achterkamp, Telemedicine Group, Roessingh Research and Development, Roessinghsbleekweg 33b, Enschede 7522AH, Overijssel, The Netherlands.
Email: r.achterkamp@rrd.nl

Introduction

A physically active lifestyle has significant positive effects on mental health condition[1] and prevention of chronic diseases such as cardiovascular disease, diabetes, and cancer.[2] A recent development regarding physical activity interventions is using mobile applications to achieve behavioural change. Many applications allow for tracking and scheduling of exercise, while only few applications aim at tracking physical activity over the day. Those that are available typically use an external sensor next to a smartphone, like Fitbit[3] and Samsung Gear Fit.[4] These types of services seem promising in the short-term.[5] However, the effectiveness can be further improved.

Traditional, non-mobile physical activity interventions that aim to improve level of physical activity in the general population,[6,7] frequently personalize, or tailor feedback based on theories and models from behavioural sciences to increase effectiveness and even optimize adherence to the intervention.[8] This specific type of tailoring – personalization of information or feedback based on an individual's score on constructs from behavioural sciences – is called *adaptation.*[9] Whereas traditional interventions frequently use adaptation of feedback, this is rarely applied in modern-day, mobile physical activity applications.[10]

A source for identifying how to apply adaptation is social cognition models (SCMs). These define the cognitive factors that underlie social patterns of behaviour. Three well-known examples are the social cognitive theory (SCT), theory of planned behaviour (TPB), and transtheoretical model (TTM).[8] The SCT assumes that motivation and action are influenced by forethought.[11] It describes three types of expectancies: situation outcome expectancy, action outcome expectancy, and perceived self-efficacy. It states that personal sense of control makes it possible to change behaviour; if people believe they can take action to accomplish a certain goal, they become more inclined to do so and feel more committed to the decision. The TPB states that behaviour is preceded by intentions, that is, motivation or plans to exert effort to perform behaviour.[12] Intentions are constituted by attitudes, subjective norms, and perceived behavioural control. By influencing the various beliefs properly, behaviour can be changed and maintained. Finally, the TTM assumes changing behaviour requires progress through five stages (Table 1) and different cognitions may be of importance at different stages.[13] The stages can be entered and exited at any point and it is possible to relapse to an earlier stage. Next to these stages, the model includes several other constructs: a decisional balance (benefits versus costs), self-efficacy (confidence that one can engage in healthy behaviour; temptation to engage in unhealthy behaviour) and processes of change (activities that people engage in to progress through the stages).

Indeed, research shows that traditional interventions that use adaptation based on constructs from SCMs, like attitudes, self-efficacy, stage of change, social support or processes of change, showed significantly larger effect sizes than interventions that did not tailor on these constructs.[9,13,14] In addition, guidelines for designing effective physical activity interventions strongly recommend tailoring feedback.[6,15] Furthermore, O'Reilly and Spruijt-Metz[16] conclude a systematic review by stating that with respect to using technology for assessment and promotion of physical activity, more research is needed on the effectiveness of interventions that combine real-time, tailored, and adaptive feedback.

Primary objective

It is hypothesized that implementing knowledge from behavioural sciences into modern-day, mobile physical activity applications can further improve their effectiveness, just as in traditional interventions. As such, the aim of this study is to investigate (1) the relation between self-efficacy

Table 1. Stages of change and their corresponding definition.

Stage of change	Definition
Precontemplation	No intention to change behaviour within 6 months
Contemplation	Intention to change behaviour within the next 6 months
Preparation	Intention to take steps to change behaviour within the next month
Action	Changed behaviour for less than 6 months
Maintenance	Changed behaviour for more than 6 months

and objectively measured level of physical activity; (2) the relation between stage of change and objectively measured level of physical activity; and (3) compare level of physical activity between patients and healthy adults.

Secondary objective

Based on the results typical users, that is, personas, will be identified. Tailored feedback strategies will be developed for these personas, which can be used to improve the effectiveness of mobile physical activity coaches in the future. The reason for choosing self-efficacy and stage of change is that these are two aspects which are of central importance in most SCMs and common in traditional interventions.

Method

Data were available for secondary analysis from previous studies performed from 2008 to 2011.[17,18] Data about stage of change, level of self-efficacy, and objectively measured physical activity were collected, but not used in any way during and after completion of these studies.

Participants

Data of 407 participants were analysed of which 82 were patients diagnosed with one of the following conditions: chronic obstructive pulmonary disease (COPD) (n = 39), chronic low back pain (CLBP) (n = 20), or cancer (n = 23). All of these patients were grouped together, as the literature shows comparable physical activity data of the separate groups.[17,18] The patient group consisted of 43 women and 39 men, averaging 60 years of age (standard deviation (SD) = 12). The healthy group consisted of 149 women and 176 men. All participants signed an informed consent. A local ethics committee reviewed and approved the study.

Equipment

Two types of mobile activity monitoring system were used: the Activity Coach (AC; see Figure 1)[5,17,18] and a Commercial Activity Monitoring Device (CAMD). The AC was worn by 139 participants (82 patients and 57 healthy controls (AC control participants)). The CAMD was worn by 268 healthy controls (CAMD control participants).

The AC consists of a sensor (MTx-w) and a smartphone (HTC). The sensor includes a tri-axial accelerometer which is used to measure physical activity. It is worm on the hip and sends data to the smartphone through a Bluetooth® connection. Op den Akker et al.[19] provide a complete description of the system.

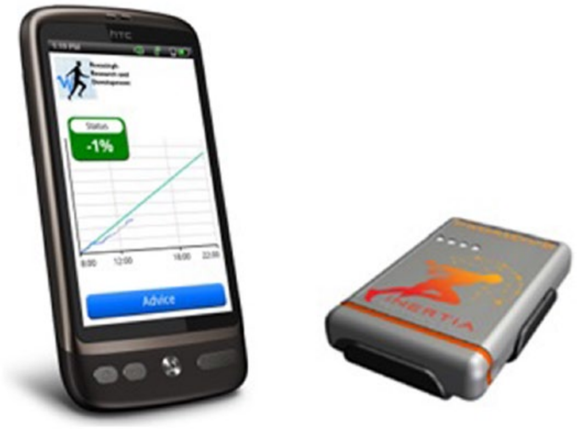


Figure 1. Activity Coach.

The CAMD consists of a tri-axial accelerometer. The dimensions of the device are approximately 3 by 3 by 1 cm. Users can wear the device in their pocket, on their belt, or as a necklace.

Procedure

Participants wore the AC the entire day, for seven consecutive days. The goal here was to obtain a baseline measurement of the users' level of physical activity. They did not receive any kind of feedback during these 7 days; only physical activity was measured throughout the day. Additionally, patients were asked to complete a questionnaire assessing their stage of change[13] and working status at the beginning of the experiment.

Participants using the CAMD completed a questionnaire assessing level of self-efficacy regarding physical activity at the start of the experiment.[20] Low, average, and high levels of self-efficacy corresponded to scores of 5 through 12, 13 through 17 and 18 through 25, respectively. Hereafter, participants wore the device the entire day, for 3 weeks, to obtain a baseline measurement of their level of physical activity. They did not receive any kind of feedback during these 3 weeks; only physical activity was measured throughout the day.

Data analysis

The accelerometer of the AC calculates activity counts per minute (CPM) as output which was processed in MATLAB to gain insight in the level of physical activity and physical activity pattern. Level of physical activity was defined as the average amount of Integral of the Modulus of the Accelerometer (IMA) counts per minute per day. A day was considered a valid measurement day if data are collected for 50% of an hour for at least 6 h per day. Furthermore, every day part should contain at least 2 h of valid data. The day parts were defined as morning (08:00 a.m.–13:00 p.m.), afternoon (13:00 p.m.–17:00 p.m.), and evening (17:00 p.m.–22:00 p.m.). The averages of IMA counts per minute per day part were calculated to investigate differences in physical activity patterns over the day.

The CAMD calculates a ratio between calorie expenditure and basic metabolism, based on age, length, weight, and sex, to estimate level of physical activity. It uses PAL as output measure, which has a minimum of 1.1. If participants show a PAL of 1.7 or above, they are considered active. The exact calculation cannot be disclosed, since the CAMD is commercially available.

Statistical analysis

The correlation between age and level of physical activity was calculated and an analysis of variance (ANOVA) was performed to examine differences between sexes regarding level of physical activity and the effect of working status on level of physical activity to identify possible confounding factors. The latter only investigated this effect for the patient group, since data regarding working status were not available for the AC control group. Patients were classified as unemployed (less than 12 h of work per week), part-time (between 12 and 36 h of work per week), or full-time (more than 36 h of work per week).

A univariate ANOVA was performed to test the difference between the level of physical activity of patients and AC control participants. With respect to the patient group, the difference in level of physical activity per stage of change was analysed using an ANOVA. Furthermore, the level of physical activity of patients per stage of change was compared to the level of physical activity of AC control participants.

Repeated measures-MANOVA was executed to analyse level of physical activity per day part (morning, afternoon, evening); testing differences in patterns between patients and AC controls, and between patients per stage of change. An ANOVA was performed to test whether CAMD control participants with different levels of self-efficacy (low, average, high) show different levels of physical activity.

Results

Results regarding the AC

The results show no significant correlation between age and average daily level of physical activity for neither the patient group (r=−.107, p=.356) nor the AC control group (r=.170, p=.21). The ANOVA indicates no significant difference in level of physical activity between sexes in the AC control group (F(1, 55)=1.99, p=.164) or in the patient group (F(1, 75)=3.34, p=.072). Regarding working status, 56 patients were unemployed, 11 had a part-time job, and 10 worked full-time. No significant difference in level of physical activity was found between working status (F(2, 74)=1.75, p=.182). Based on these results, it can be assumed that level of physical activity was not influenced by age, sex, or working status in the current study.

The univariate ANOVA shows that patients (mean IMA=947.77) are significantly less active than AC controls (mean IMA=1089.6) (F(1, 132)=8.58, p=.004). Within the patient group, there is a significant difference in level of physical activity per stage of change (F(3, 72)=4.00, p=.011) (Figure 2). Patients in the contemplation, preparation, and action stage of change are significantly less active than patients in the maintenance stage of change (β=−197.69 (t=−1.99, p=.051); β=−215,69 (t=−3.03, p=.003); and β=−221.67 (t=−2.01, p=.048), respectively).

Results also show a significant difference in level of physical activity between patients per stage of change and control participants (F(4, 128)=5.15, p=.001). Contrasts show that patients in the contemplation, preparation and action stage of change are less active than AC controls (β=−226.17 (t=−2.35, p=.020); β=−244,26 (t=−3.69, p<.001); and β=−250.25 (t=−2.33, p=.021), respectively). No significant difference was found in level of physical activity between patients in the maintenance stage of change and AC control participants (β=−28.58 (t=−.51, p=.61)) (Figure 2).

Regarding physical activity pattern, the repeated-measures-MANOVA shows a significant difference in activity per day part (W=.77, p<.001) (GG: F(1.63, 198.77)=28.57, p<.001); physical activity over the day of all participants combined shows a quadratic trend from morning to evening (F(1, 122)=11.93, p=.001). The interaction effect between activity per day part and group (patient/

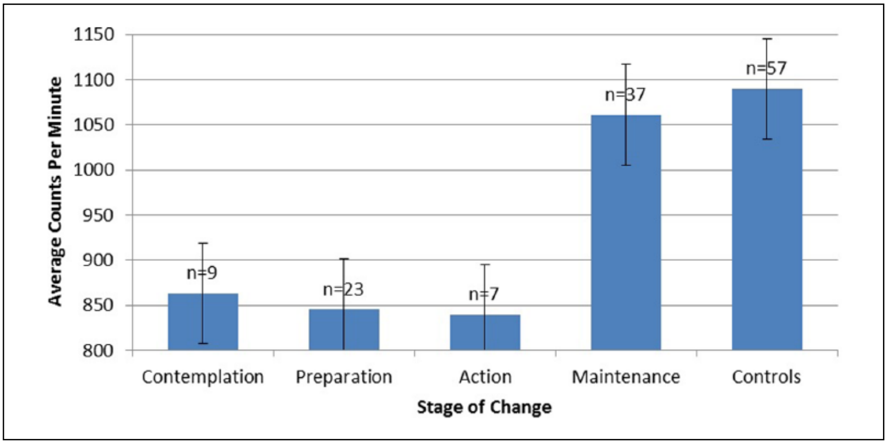


Figure 2. Average CPM per stage of change for patients compared to the average CPM of AC control subjects.

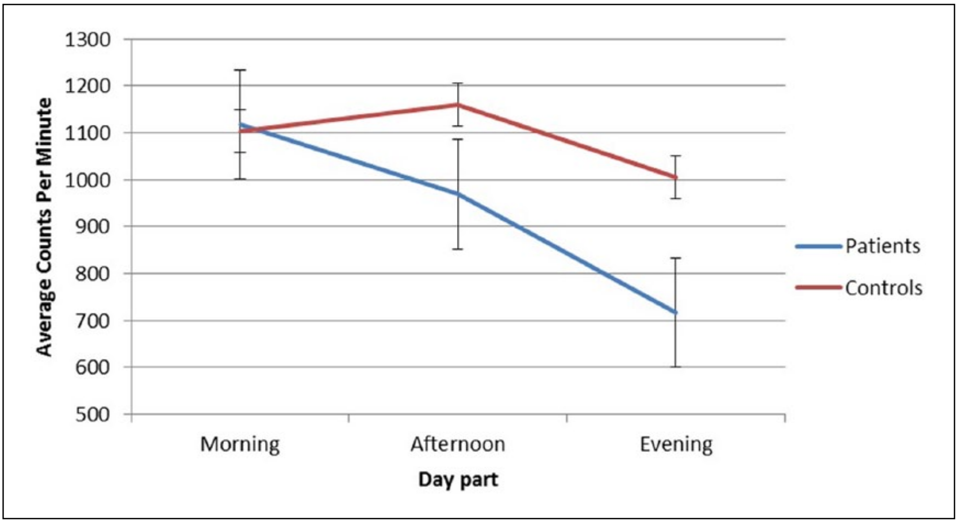


Figure 3. Average CPM per day part for patients and AC control subjects.

AC control) is significant (GG: F(1.63, 198.77)=9.45, p<.001), indicating the difference per day part is different for patients than for AC controls. Figure 3 shows that the decline in level of physical activity over the day is much steeper for patients than for AC controls; they are as active as AC controls in the morning (β=13.042 (t=.17, p=.865)), but whereas AC controls show an increase of physical activity in the afternoon, patients show a decrease (β=−191,489 (t=−3.27, p=.001), and an even steeper decrease than AC controls in the evening (β=−287.064 (t=−4.95, p<.001)).

With respect to the group of patients, the difference in activity per day part is not different per stage of change (W=.700, p<.001) (GG: F(4.61, 95.35)=1.15, p=.34). To provide an overview, Figure 4 shows the level of physical activity per day part for patients per stage of change as compared to the level of physical activity per day part of AC control participants. Whereas AC control

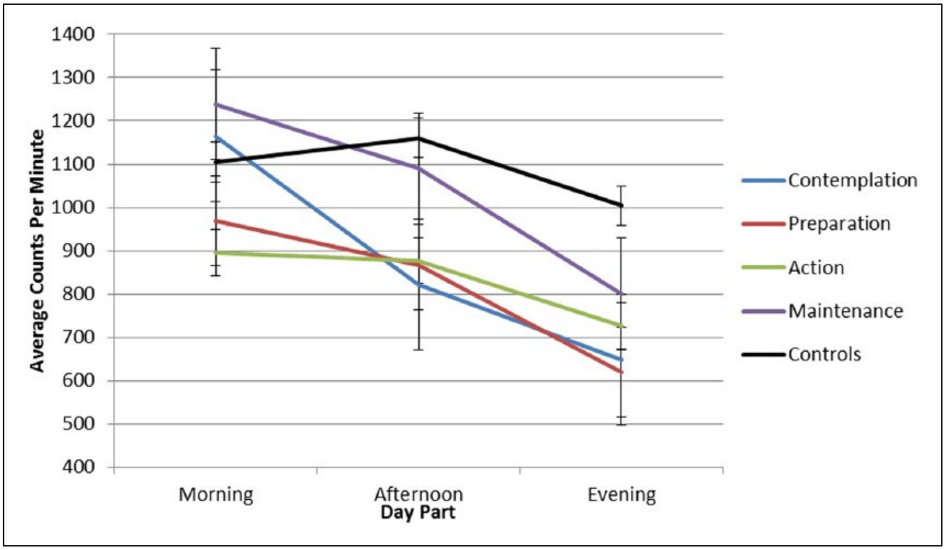


Figure 4. Average CPM per day part for patients per stage of change as compared to AC control subjects.

participants show a small drop in level of physical activity over the day, all patients show the same pattern of high decline in level of physical activity from morning till evening, regardless of the participant's stage of change.

Results regarding the CAMD

With respect to the CAMD and the relationship between self-efficacy and physical activity, sex was added to the model as a fixed factor, as the ANOVA showed a significant difference in level of physical activity between sexes (F(1, 266) = 6.55, p = .011); men (mean PAL = 1.657; SD = .133) are more active than women (mean PAL = 1.616; SD = .124).

Most CAMD participants were classified as having an average level of self-efficacy regarding physical activity (n = 144), 60 participants reported a high level of self-efficacy and 55 participants indicated a low level of self-efficacy. The test shows a significant difference in level of physical activity per category of self-efficacy (F(2, 253) = 8.69, p < .001). The interaction effect with sex is not significant. Contrasts indicate that participants with a low or average level of self-efficacy are significantly less active than participants with a high level of self-efficacy (β = −.080 (t = −2.07, p = .039) and β = −0.090 (t = −2.70, p = .007), respectively) (Figure 5).

Discussion

The primary aim of this study was to investigate (1) the relation between self-efficacy and objectively measured level of physical activity, (2) the relation between stage of change and objectively measured level of physical activity, and (3) compare level of physical activity between patients and healthy adults, in mobile physical activity interventions. Secondary, based on the results, typical users were identified and corresponding tailored feedback strategies were developed. Results show

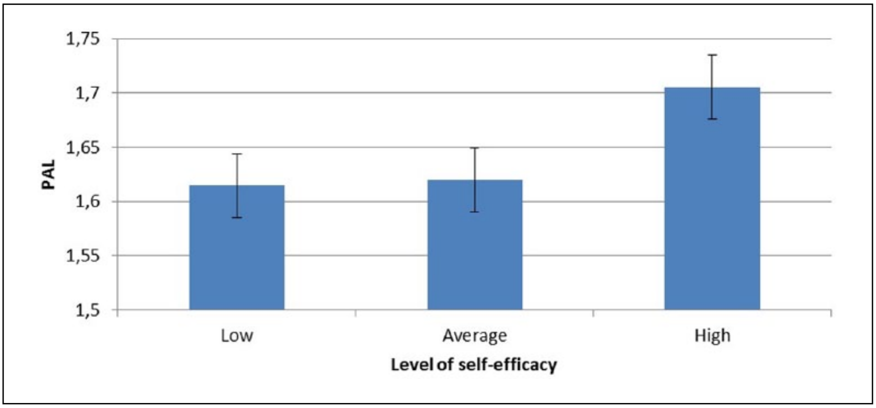


Figure 5. Average level of physical activity per category of self-efficacy.

that the three factors are significantly related to objectively measured physical activity: self-efficacy, stage of change and being healthy or suffering from a disease.

With respect to self-efficacy, higher levels of self-efficacy are related to higher levels of physical activity. The more participants believe that being sufficiently physically active is within their control, the higher their level of physical activity. These findings are consistent with traditional physical activity research, which shows that participants who have not started to exercise regularly show low levels of self-efficacy, whereas those who have started show high levels of self-efficacy.[21]

Having a chronic disease also influences level of physical activity. Patients are less active and show a steeper decline in level of physical activity over the day than healthy participants. Research suggests that patients tend to do all must-tasks (e.g. cleaning, groceries) in the morning, leaving them with little energy to do social and fun activities in the evening.[5,17]

With respect to stage of change, patients in the maintenance stage of change are more active than patients in other stages of change; they are as active as healthy participants. However, patients in the maintenance stage of change show an equally large drop in level of physical activity over the day as other patients and, as such, have an improper activity pattern.

Based on these results, participants can be categorized into eight typical personas, who should receive different coaching strategies based upon the three important variables stage of change, self-efficacy and level of physical activity (Tables 2 and 3).

Based on stage of change, participants can be categorizes as either having (contemplation, preparation, action) or not having (precontemplation, maintenance) an intention to change behaviour. Based on the activity pattern, participants can show a proper or improper level of physical activity. A proper level of physical activity means sufficient physical activity and a balanced physical activity pattern; improper indicates insufficient physical activity or an imbalanced pattern. Regarding self-efficacy, participants can be categorized as having ‘low to average’ or ‘high’ self-efficacy. Low and average levels of self-efficacy were taken together, as these participants did not show differences in level of physical activity. Ideally, scores on these constructs should be assessed regularly to identify whether they are still categorized as the correct persona, or if they have changed to, for example, a higher level of self-efficacy, for which an adjustment of the coaching strategy is needed.

The personas described above can be used to develop corresponding feedback strategies that can be included into new mobile physical activity applications. It is clear that coaching should at least be tailored to users’ level of self-efficacy, stage of change and physical activity pattern. As

Table 2. Personas with intention to change.

Self-efficacy	Level of activity	
	Proper	Improper
Low–average	Persona 1	Persona 2
High	Persona 3	Persona 4

Table 3. Personas without intention to change.

Self-efficacy	Level of activity	
	Proper	Improper
Low–average	Persona 5	Persona 6
High	Persona 7	Persona 8

high self-efficacy not only increases intention, but also leads to actual performance of the target behaviour,[22] much research has focused on how self-efficacy can be influenced and especially on how to increase it. Bandura[23] describes four sources that can be used to increase self-efficacy: mastery experience, vicarious experience, social persuasion and physiological and emotional states. Regarding personas 1, 2, 5, and 6, who have low levels of self-efficacy, mastery experience could be implemented by setting challenging but attainable, personalized goals,[24] leading to success experiences. Adding optional data sharing leads to vicarious experience and additionally sending persuasive feedback messages makes for higher exerted effort of users to achieve their goal. A meta-analysis showed that of these four strategies to increase self-efficacy, feedback on previous performance or previous performance of similar others cause the highest effect sizes, followed by vicarious experience.[25] As such, this might also be hypothesized to be the most effective strategy to include in mobile physical activity applications.

Regarding stage of change, 10 specific strategies to move from stage to stage, or processes of change, have received much attention and empirical support.[26] Five can be identified as experiential processes (increasing awareness, emotional arousal, social reappraisal, social liberation, and self-reappraisal), and the other five are referred to as behavioural processes (stimulus control, social support, counter conditioning, rewarding, committing). Experiential processes are primarily used for early stages, while behavioural processes are recommended for later stages.[23] Therefore, coaching for personas 1, 2, 3, and 4 should focus on behavioural processes of change, whereas coaching for personas 5, 6, 7, and 8 should focus on experiential processes.

The coaching strategies were implemented into the AC (Figure 1) and are currently tested in a field study. First, level of self-efficacy, stage of change and level of physical activity are assessed at baseline, after which participants are automatically identified as one of the eight personas, which determines what feedback messages they will receive during the intervention; different personas receive different feedback messages.

Conclusion

Just as traditional physical activity interventions, modern-day mobile physical activity applications should include adaptation and tailored feedback strategies into their coaching, which might lead to increased effectiveness and hopefully to even better intervention adherence, and adherence to physical activity guidelines. This is not yet known. However, this study can be regarded as first step towards testing this. It identifies personas and provides guidelines for development of feedback that takes into account individual scores on constructs from behavioural sciences. The next step is to test these findings in daily life. Additionally, there are many other factors associated with physical activity (e.g. social support, benefits, barriers, etc.), and as such, future research should investigate further adaptation and tailoring of feedback strategies in mobile physical activity interventions using knowledge from social cognition models.

Declaration of conflicting interests

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding

This publication was supported by the Dutch national program COMMIT (project P7 SWELL).

References

1. Jonsdottir IH, Rödjer L, Hadzibajramovic E, et al. A prospective study of leisure-time physical activity and mental health in Swedish health care workers and social insurance officers. *Prev Med* 2010; 51(5): 373–377.
2. Warburton DER, Nicol CW and Bredin SSD. Health benefits of physical activity: the evidence. *Can Med Ass J* 2006; 174(6): 801–809.
3. Fitbit. San Francisco, CA: Fitbit, 2015, http://www.fitbit.com (accessed 13 June 2015).
4. Samsung Gear Fit. Seoul, South Korea, 2015, http://www.samsung.com/global/microsite/gear/index.html (accessed 13 June 2015).
5. Van Weering MGH. *Towards a new treatment for chronic low back pain patients: using activity monitoring and personalized feedback*. Doctorial Thesis, Roessingh Research and Development, Enschede, 2011.
6. Dishman RK and Buckworth J. Increasing physical activity: a quantitative synthesis. *Med Sci Sports Exerc* 1996; 28(6): 706–719.
7. Marcus BH, Bock BC, Pinto BM, et al. Efficacy of an individualized, motivationally tailored physical activity intervention. *Ann Behav Med* 1998; 20(3): 174–180.
8. Conner M and Norman P. *Predicting health behaviour*. 2nd ed. Berkshire: Open University Press, 2005.
9. Hawkins R, Kreuter M, Resnicow K, et al. Understanding tailoring in communicating about health. *Health Educ Res* 2008; 23(3): 454–466.
10. Op den Akker H, Jones VM and Hermens H. Tailoring real-time physical activity coaching systems: a literature survey and model. *User Model User: Adap* 2014; 24: 351–392.
11. Bandura A. Self-efficacy mechanism in human agency. *Am Psychol* 1982; 37: 122–147.
12. Ajzen I. The theory of planned behaviour. *Organ Behav Hum Dec* 1991; 50: 179–211.
13. Prochaska JO and DiClemente CC. Stages and processes of self-change of smoking: toward an integrative model of change. *J Consult Clin Psychol* 1983; 51: 390–395.
14. Noar S, Benac C and Harris M. Does tailoring matter? A meta-analytic review of tailored print health behaviour change interventions. *Psychol Bull* 2007; 133(4): 673–693.
15. Greaves CJ, Sheppard KE, Abraham C, et al. Systematic review of reviews of intervention components associated with increased effectiveness in dietary and physical activity interventions. *Public Health* 2011; 11(119): 1471–2458.
16. O'Reilly GA and Spruijt-Metz D. Current mHealth technologies for physical activity assessment and promotion. *Am J Prev Med* 2013; 45(4): 501–507.
17. Tabak M, Vollenbroek-Hutten MMR, van der Valk PDLPM, et al. Telemonitoring of daily activity and symptom behavior in patients with COPD. *Int J Telemed Appl* 2012; article ID 438736.

18. Dekker-van Weering MGH, Vollenbroek-Hutten MMR and Hermens HJ. Do personalized messages about activity patterns stimulate patients with chronic low back pain to change their activity behavior on a short term notice? *Appl Psychophysiol Biofeedback* 2012; 37(2): 81–89.
19. Op den Akker H, Tabak M, Marin-Perianu M, et al. Development and evaluation of a sensor-based system for remote monitoring and treatment of chronic diseases – the continuous care & coaching platform. In: *6th international symposium on eHealth services and technologies*, Geneva, 3–4 July, 2012.
20. Rodgers WM, Wilson PM, Hall CR, et al. Evidence for a multidimensional self-efficacy for exercise scale. *Res Q Exerc Sport* 2008; 79(2): 222–234.
21. Marcus BH, Selby VC, Niaura RS, et al. Self-efficacy and the stages of exercise behaviour change. *Res Q Exerc Sport* 1992; 61(1): 60–66.
22. Gist ME and Mitchell TR. Self-efficacy: a theoretical analysis of its determinants and malleability. *Acad Manage Rev* 1992; 17(2): 183–211.
23. Bandura A. Self-efficacy. In: Ramachaudran VS (ed.) *Encyclopedia of human behaviour*. New York: Academic Press, 1994, pp. 71–81.
24. Locke EA and Latham GP. Building a practically useful theory of goal setting and task motivation: a 35-year odyssey. *Am Psychol* 2002; 57(9): 705–717.
25. Ashford S, Edmunds J and French DP. What is the best way to change self-efficacy to promote lifestyle and recreational physical activity? A systematic review with meta-analysis. *Br J Health Psychol* 2010; 15: 265–288.
26. Velicer WF, Prochaska JO, Fava JL, et al. Smoking cessation and stress management: applications of the transtheoretical model of behaviour change. *Homeostasis Hlth Dis* 1998; 38: 216–233.

Article

Health Informatics Journal

Health Informatics Journal
2018, Vol. 24(1) 103–110
© The Author(s) 2016
Reprints and permissions:
sagepub.co.uk/journalsPermissions.nav
DOI: 10.1177/1460458216654514
journals.sagepub.com/home/jhi
SAGE

YouTube®: An ally or an enemy in the promotion of living donor kidney transplantation?

Fabrizio Bert, Maria Rosaria Gualano, Gitana Scozzari, Marta Alesina, Antonio Amoroso and Roberta Siliquini
University of Turin, Italy

Abstract

The aim of the study is to evaluate the availability and accuracy of the existing Italian-language medical information about living donor kidney transplantation on YouTube®. For each video, several data were collected, and each video was classified as "useful," "moderately useful" and "not useful." Globally, the search resulted in 306 videos: 260 were excluded and 46 included in the analysis. The main message conveyed by the video was positive in 28 cases (60.9%), neutral in 16 (34.8%) and negative in 2 (4.4%). The mean amount of visualizations was 3103.5 (range: 17–90,133) and the mean amount of "likes" 2.7 (range: 0–28). Seven videos (15.2%) were classified as "useful," 21 (45.7%) as "moderately useful" and 18 (39.1%) as "not useful." This study showed that a very few videos in Italian about living donor kidney transplantation are available on YouTube, with only 15 percent of them containing useful information for the general population.

Keywords

e-health, health information on the Web, health promotion, living donor kidney transplantation, YouTube®

Introduction

To date, it is well known that kidney transplantation improves both the quantity and quality of life in end-stage renal disease patients compared with hemodialysis,[1] and living donor kidney transplantation (LDKT) shows increasingly better results compared to cadaveric donors. LDKT, indeed, has increased the number of organs available for transplantation, thus shortening the relevant waiting lists and has improved the post-transplant outcomes by reducing graft failure and extending graft survival.[2] Furthermore, LDKT provides a greater assurance of transplantation before dialysis than the deceased organ donation, thus reducing dialysis-related complications and costs,[3] and gives the opportunity to plan the surgical procedure, thus permitting to optimize the recipient's

Corresponding author:
Maria Rosaria Gualano, Department of Public Health Sciences, University of Turin, via Santena 5bis, 10126 Turin, Italy.
Email: mariarosaria.gualano@unito.it

conditions. Moreover, LDKT has shown to result in better patient and allograft outcomes compared to deceased donor transplant,[4] and longitudinal studies report lower perioperative mortality rates and no renal deterioration in donors.[5]

In Italy, approximately 1700 kidney transplantations are performed annually, with less than 15 percent of LDKT.[6] Moreover, considering that the current national waiting list includes more than 6500 end-stage kidney disease patients, there is a clear organ shortage crisis.[6] Since the little number of LDKT performed could be also related to a lack of knowledge among the general population,[7,8] sharing information and awareness on this topic is a critical goal for the Public Health community.

To date, the Internet represents the largest and most-widely used source of medical information among the general population and chronic disease patients.[9] YouTube® is a video-sharing web service that has rapidly become a popular Internet-based mass media, also in the field of health-related contents. Previous studies have documented the availability on YouTube of health-related videos on different topics such as immunization,[10] human papilloma virus vaccination,[11] dialysis[12] and organ donation,[13] but no previous study has analyzed the available YouTube contents on LKDT.

The aim of this study is to evaluate the availability and accuracy of the existing Italian-language medical information about LDKT on YouTube and to analyze the user's feedback on the contents and quality of such videos.

Methods

A web search was performed on the YouTube web page (www.youtube.com) on 25 June 2015. The searched keywords were "transplantation," "donation," "donors," "kidney" and "living." The results of the search were sorted by relevance, which is the default setting of the website. Videos in Italian language only were included. Some videos were excluded for lack of relevance with the aim of the study and for not being in Italian. Duplicated videos were also excluded.

Videos were evaluated in duplicate independently by two authors (G.S. and M.A.), resident doctors in Public Health with experience in health education and health promotion, and disagreements were resolved by discussion with a third author (F.B.), university researcher in Public Health with experience in health promotion and e-health. Videos' evaluation was discussed with a fourth author (A.A.), full professor of Medical Genetics and chief director of the Regional Centre for Transplantations. Data were entered into an electronic Excel® spreadsheet designed ad hoc for the purposes of this study, including all data collected.

For each video, the following data were collected: video title, URL, upload source, upload date, video length, number of views, number of "likes" and "dislikes," along with type of video (classified according to the main aim of the video, as follows: educational video, medical video, surgical technique video, journalism, personal experience and so on), target audience and main message conveyed. Similar to Keelan et al.,[10] the main message was classified as *positive* if the video supported LDKT, portraying it positively (e.g. described LDKT as a social good or described the benefits and safety of LDKT), as *negative* if the video portrayed LDKT negatively (e.g. emphasized the risk of complications, advocated against organ donation, supported theories of conspiracy or collusion between supporters of LDKT and the organ illegal commerce) and as *neutral* when it was not possible to identify a message since the video contained a debate or was nonpartisan.

Each video was then evaluated for the presence or absence of information on 13 critical content domains: the presence of institutional or hospital address and reference, to provide a reference point for patients and their relatives; pre-transplant donors' medical evaluation description; pre-transplant recipients' medical evaluation description; living donor transplantation advantages over cadavers donation; possibility of pre-dialysis transplantation; surgical technique; possibility of laparoscopic nephrectomy; presence of an interview of a physician, of another health-care worker, of a donor or a recipient; legislation rules; and possibility of Samaritan donation. Depending on the weighting of the 13 items, these were given different scores (from 0.5 to 2 points); each item was scored on the basis of the presence in the video, that is, not mentioned (zero points), mentioned briefly (half score) or mentioned in detail (total score). By calculating the total score, each video was classified as "useful" (>5 points), "moderately useful" (>2–≤5 points) and "not useful" (≤2 points).

Statistical analysis

Descriptive statistical analysis was performed using frequencies and percentages for the categorical variables, mean with standard deviation and minimum–maximum values for quantitative variables. In order to evaluate the different video characteristics between groups (useful, moderately useful and not useful videos), the chi-square test was used, with Fisher's correction when needed, for categorical variables. For continuous measures, one-way analysis of variance models were used to determine relationships with video groups. The statistical significance level was set at $p \leq 0.05$. Statistical analyses were performed with STATA 11. No ethical approval was required for this study.

Results

The search globally resulted in 306 videos. Of them, 260 were excluded because relating to cadaveric organ transplantation (n = 74) or transplantation of organs other than kidney or organ transplantation in general (n = 83), because they were in languages other than Italian (n = 3) or for being completely inappropriate for the purpose of the study, that is, relating to topic unrelated to kidney transplantation (n = 100). Thus, 46 videos were included in the analysis (Figure 1).

Videos had been uploaded between December 2007 and October 2015, 25 of them (54.3%) between 2013 and 2015. The mean video length was 574 ± 717 s (range: 56–3431 s); using the length categories of YouTube, there were 17 short videos (i.e. <240 s), 22 medium videos (i.e. 240–1200 s) and 7 long videos (i.e. >1200 s).

Videos had been uploaded by a local or web-based television channel in 21 cases (45.7%), by a national TV channel in 6 (13.0%; 5 videos by the Italian National public TV channel and 1 by a private television), by a university in 5 cases (10.9%), by an individual in 6 (13.0%), by an hospital in 4 (8.7%) and by a scientific society in 4 (8.7%). The videos were classified as journalism in 23 cases (50%), surgical technique documentation in 7 (15.2%), personal experience in 6 (13.0%), educational video in 5 (10.9%), medical or scientific in 1 (2.2%) and "other" in 4 (8.7%; more specifically, a politic message in 1 case and a computer game in 3 cases). The main source of information in the mentioned videos was classified as medical in 30 cases (65.2%), patients' personal experience in 6 (13.0%), journalism in 5 (10.9%) and as "no recognizable source" in 5 (10.9%). The most common target of the videos was the general population in 36 cases (78.3%), physicians and health-care professionals in 7 cases (15.2%) and adolescents and young adults in 3 cases (6.5%). The main message conveyed by the video was classified as positive in 28 cases (60.9%), neutral in 16 (34.8%) and negative in 2 (4.4%). With regard to the viewers' appreciation of the videos, the mean amount of visualizations was 3103.5 ± 13,270.3 (range: 17–90,133), with 34 videos (73.9%) having less than 1000 visualizations. The mean number of "likes" per video was 2.7 ± 5.1 (range: 0–28) while for "dislikes" was 0.5 ± 1.5 (range: 0–9).

The usefulness criteria of contents are shown in Figure 2. Mean usefulness score was 2.9 ± 2.2. Globally, 7 videos (15.2%) were classified as useful, 21 (45.7%) as moderately useful and 18 (39.1%) as not useful.

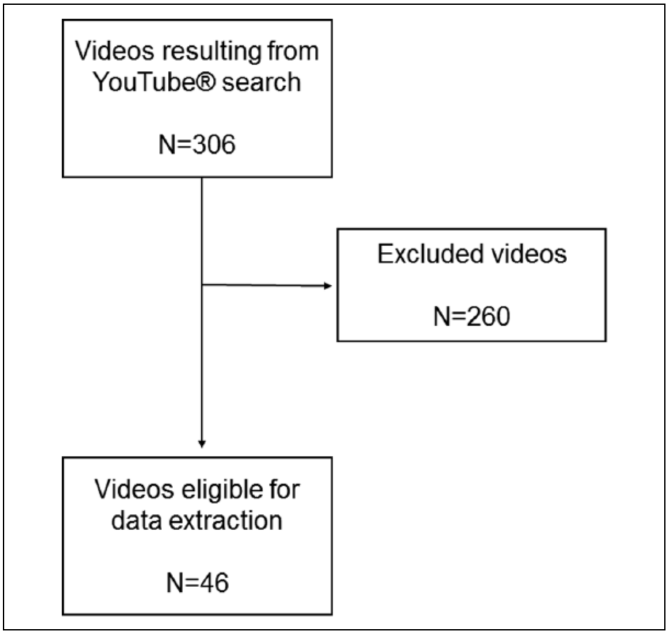

Videos resulting from YouTube® search
N=306
Excluded videos
N=260
Videos eligible for data extraction
N=46

Figure 1. Flowchart detailing the video selection process.

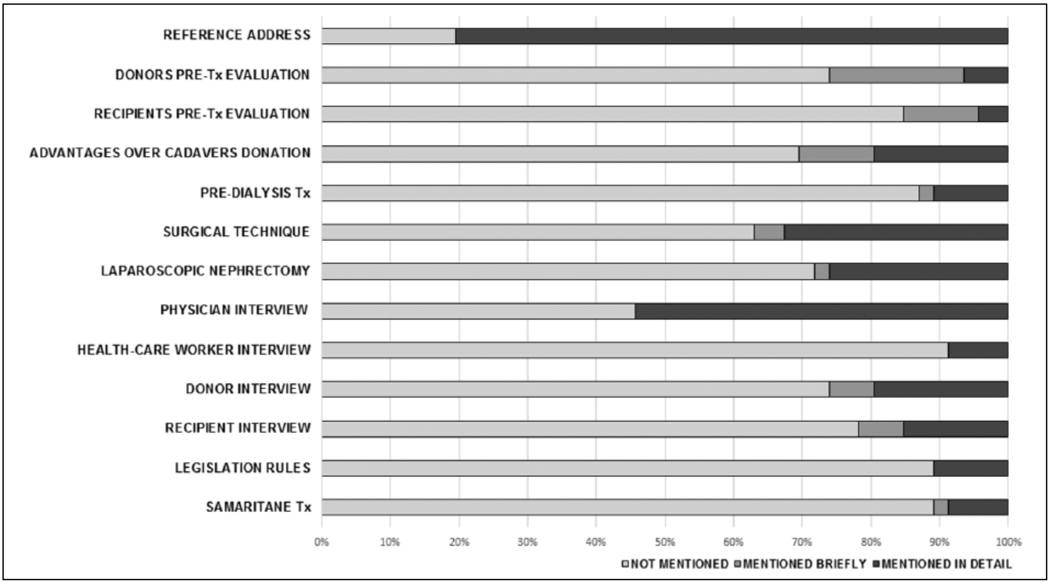


Figure 2. Usefulness contents. Data are expressed as absolute number (N = 46).
Tx: transplantation.

The three categories of videos (useful, moderately useful and not useful) showed statistically significant differences in the length of video, the uploading source, type of video, video target and main message, as reported in Table 1.

Table 1. Usefulness videos' content among video categories.

Contents	All videos (N = 46)	Useful videos (N = 7)	Moderately useful videos (N = 21)	Not useful videos (N = 18)	p value*
Length (s)					
Mean	574.2	1446.6	345.2	502.2	**0.0008**
Likes (n)					
Mean	2.7	5.1	1.3	3.4	0.178
Dislikes (n)					
Mean	0.5	0.9	0.1	0.9	0.200
Visualizations (n)					
Mean	3103.5	2411.9	1077.1	5736.7	0.555
Producer					
University	5 (10.9%)	3 (72.9%)	1 (4.8%)	1 (5.6%)	**<0.001**
Hospital	4 (8.7%)	0	0	4 (22.2%)	
Scientific society	4 (8.7%)	1 (14.3%)	2 (9.5%)	1 (5.6%)	
Local or web TV	21 (45.7%)	2 (28.6%)	15 (71.4%)	4 (22.2%)	
National TV	6 (13.0%)	1 (14.3%)	3 (14.3%)	2 (11.1%)	
Individuals	6 (13.0%)	0	0	6 (33.3%)	
Video type					
Educational	5 (10.9%)	4 (57.1%)	1 (4.8%)	0	**<0.001**
Medical	1 (2.2%)	0	1 (4.8%)	0	
Surgical technique	7 (15.2%)	0	0	7 (38.9%)	
Journalistic service	23 (50.0%)	3 (42.9%)	16 (76.2%)	4 (22.2%)	
Personal experience	6 (13.0%)	0	3 (14.3%)	3 (16.7%)	
Other	4 (8.7%)	0		4 (22.2%)	
Target					
General population	36 (78.3%)	7 (100%)	2 (95.2%)	9 (50.0%)	**0.005**
Physicians	7 (15.2%)	0	1 (4.8%)	6 (33.3%)	
Youths	3 (6.5%)	0	0	3 (16.7%)	
Source					
Medical	30 (65.2%)	7 (100%)	15 (71.4%)	8 (44.4%)	0.109
Patients	6 (13.0%)	0	2 (9.5%)	3 (16.7%)	
Journalists	5 (10.9%)	0	4 (19.1%)	2 (11.1%)	
No sources	5 (10.9%)	0	0	5 (27.8%)	
Main message					
Positive	28 (60.9%)	7 (100%)	19 (90.5%)	2 (11.1%)	**<0.001**
Negative	2 (4.4%)	0	0	2 (11.1%)	
Neutral	16 (34.8%)	0	2 (9.5%)	14 (77.8%)	

Data are expressed as mean or number (%). Statistically significant results ($p \leqslant 0.05$) are reported in bold.
*Chi-square test with Fisher's correction when needed or one-way analysis of variance models.

Discussion

In this study, we aimed not only to evaluate the availability of Italian-language videos about LDKT on YouTube but also to analyze the user's feedback on such videos. Unfortunately, we found not only a scarce amount videos but also a small number of median views per video, thus suggesting a poor interest in this topic. The lack of interest mirrors the fact that despite the well-known advantages of LDKT, it still represents a small portion of kidney transplants: 36 percent of the kidney transplantations performed in 2009 in the United States[14] and less than 15 percent of 1699 performed in Italy in 2014.[6] The reasons for the underuse of LDKT are various, including both patients' lack of knowledge about living transplantation and difficulty to identify willing and eligible donors.[7,8]

Several studies have shown that potential kidney transplant recipients tend to be reluctant to ask relatives for an organ,[15,16] and this reluctance might represents actually the main barrier to LDKT.[17–19] Moreover, most potential donors simply do not know the possibility of living donation: a study based on 78 kidney donors pointed out that only 34 percent of them already knew LDKT before the recipient was diagnosed with kidney disease, while 53 percent of them when the recipient was already in a state of advanced chronic kidney disease.[20]

Therefore, it seems to be evident that the need exists to better and properly inform both the potential recipients and the potential donors about LDKT. The use of mass media and educational videos in this topic has been previously studied. Schweitzer et al.[21] first reported in 1997 an increase in living donor rates through a family education program, and Connelly et al.[22] in 1999 demonstrated an increase in donation rates among potential donors exposed to a video featuring living donors and recipients. Alvaro et al.[19] focused on kidney donation among Hispanics by means of 30-s television ads and 60-s radio ads, demonstrating that post-intervention group showed significantly more favorable behavioral intentions to become kidney donors than control group. More recently, an educational intervention based on a 5-min iPod video proved to increase the participants willing to donate organs.[23]

Among mass media, the Internet represents to date the largest and most-widely used source for medical information: in North America, 74 percent of adult population use Internet daily, with 80 percent of users searching for health-related information,[24] while in Europe, Andreassen et al.[25] reported a rate of 71 percent of health-related seeking in Internet users. In Italy, a multicenter survey highlighted that 65 percent of respondents used Internet, with 57 percent of them using it to search health-related information.[9]

YouTube is a free web service where subscribers can upload videos and share them with a very large number of viewers. Although it represents a leading audiovisual information center of medical-related videos, its freely accessible nature allows the dissemination of misleading information:[26,27] since videos are uploaded on YouTube with no quality control, health-related videos can contain erroneous and even harmful information.[28]

Previous studies have documented the availability on YouTube of health-related videos on different topics such as immunization,[10] human papilloma virus vaccination,[11] basic life support,[26] dialysis[12] and organ donation.[13] Unfortunately, the authors have reported that misleading videos represent an important portion and that they frequently generate more interest compared to scientifically accurate videos.[10,29,30] In 2007, the first study by Keelan et al.[10] about immunization on YouTube found 48 percent of videos classified as positive, compared to 32 percent negative and 20 percent ambiguous; in that study, negative videos had a higher mean star rating and more views compared to positive videos. Lee et al.[31] found 56.5 percent of videos regarding gallstone disease to be misleading, and Steinberg et al.[29] found a fair or poor information content in 73 percent of videos focused on prostate cancer. Also, YouTube videos have been reported to show more pro-smoking than anti-smoking content[27] and to portray negatively the papilloma virus vaccination in 25 percent of cases.[11] Garg et al.[12] found only 58 percent of 115 videos on dialysis to be useful and reported that useful videos showed lower viewership per day compared to misleading videos. However, organ donation seems to be positively portrayed on YouTube: Tian[13] reported that 96 percent of the videos were positively framed for organ donation, and Chen et al.,[24] by analyzing videos about heart transplantation founded 63 percent of those containing useful information.

In this study, we found only 15 percent of videos to be useful, that is, to contain clear and correct information about LDKT, thus suggesting an important need of knowledge sharing among the general population. In our results, we found that useful videos were significantly different compared to moderately and not useful videos: the first ones were frequently uploaded by a university or a scientific society, had a medical main source and conveyed a positive message in all cases, while moderately and not useful videos were uploaded by different subjects and frequently portrayed a neutral or even negative message about LDKT. More specifically, the two videos with a negative message pointed out the phenomenon of the organ illegal commerce that may be misleading for the general population.

The principal limit of this study is related to the main feature of YouTube, which is a continuously changing source of information, deeply depending on the research time and date. Thus, since our study is based on the YouTube material available on a given date, it should be considered as a snapshot of the available information on YouTube. Furthermore, this study is limited to the videos in Italian, thus providing a national point of view.

In summary, this study showed that a very few videos in Italian about LDKT are available on YouTube, with 15 percent of them containing useful information for the general population. Since the gap between demand and supply for transplantation organs continues to grow, mass media might be very useful in promoting organ donation. The advantages of the web-based video interventions include low cost, brevity and capacity of real-time updates; thus, YouTube could be used as a very effective resource to share information with minimal costs. Since to support the promotion of the correct use of e-health is a critical function of Public Health,[9] the results of this study encourage to more carefully monitor the information on LDKT conveyed by YouTube videos and to support the production and sharing of high-quality videos.

Acknowledgements

This article has been presented as oral presentation at the 48th National Congress of the Italian Society of Hygiene, Preventive Medicine and Public Health (October 2015, Milan, Italy). This article has not been submitted for publication nor has it been published in whole or in part elsewhere. The research does not involve human participants or animals, and no informed consent was needed.

Declaration of conflicting interests

The author(s) declared no potential conflicts of interest with respect to the research, authorship and/or publication of this article.

Funding

The author(s) received no financial support for the research, authorship and/or publication of this article.

References

1. Wolfe RA, Ashby VB, Milford EL, et al. Comparison of mortality in all patients on dialysis, patients on dialysis awaiting transplantation, and recipients of a first cadaveric transplant. *N Engl J Med* 1999; 341: 1725–1730.
2. Davis CL and Demonico FL. Living-donor kidney transplantation: a review of the current practices for the live donor. *J Am Soc Nephrol* 2005; 16: 2098–2110.
3. Friedewald JJ and Reese PP. The kidney-first initiative: what is the current status of preemptive transplantation? *Adv Chronic Kidney Dis* 2012; 19: 252–256.
4. Terasaki PI, Cecka JM, Gjertson DW, et al. High survival rates of kidney transplants from spousal and living unrelated donors. *N Engl J Med* 1995; 333: 333–336.
5. Najarian JS, Chavers BM, McHugh LE, et al. 20 years or more of follow-up of living kidney donors. *Lancet* 1992; 340: 807–810.
6. www.salute.gov.it (accessed 17 July 2015) (Italian Minister of Health web page).
7. Boulware LE, Hill-Briggs F, Kraus ES, et al. Effectiveness of educational and social worker interventions to activate patients' discussion and pursuit of preemptive Living Donor Kidney Transplantation: a randomized controlled trial. *Am J Kidney Dis* 2013; 61: 476–486.

8. Kranenburg LW, Zuidema WC, Weimar W, et al. Psychological barriers for living kidney donation: how to inform the potential donors? *Transplantation* 2007; 84: 965–971.
9. Siliquini R, Ceruti M, Lovato E, et al. Surfing the internet for health information: an Italian survey on use and population choices. *BMC Med Inform Decis Mak* 2011; 11: 21.
10. Keelan J, Pavri-Garcia V, Tomlinson G, et al. YouTube as a source of information on immunization: a content analysis. *JAMA* 2007; 298: 2482–2484.
11. Ache K and Wallace L. Human papillomavirus vaccination coverage on YouTube. *Am J Prev Med* 2008; 35: 389–392.
12. Garg N, Venkatraman A, Pandey A, et al. YouTube as a source of information on dialysis: a content analysis. *Nephrology* 2015; 20: 315–320.
13. Tian Y. Organ donation on Web 2.0: content and audience analysis of organ donation videos on YouTube. *Health Commun* 2010; 25: 238–246.
14. www.srtr.org (accessed 17 July 2015) (Scientific Registry of Transplant Recipients web page).
15. Waterman AD, Stanley SL, Covelli T, et al. Living donation decision making: recipients' concerns and educational needs. *Prog Transplant* 2006; 16: 17–23.
16. Pradel FG, Limcangco MR, Mullins CD, et al. Patients' attitudes about living donor transplantation and living donor nephrectomy. *Am J Kidney Dis* 2003; 41: 849–858.
17. Pradel FG, Suwannaprom P, Mullins CD, et al. Short-term impact of an educational program promoting live donor kidney transplantation in dialysis centers. *Prog Transplant* 2008; 18: 263–272.
18. Rudow DL, Chariton M, Sanchez C, et al. Kidney and liver living donors: a comparison of experiences. *Prog Transplant* 2005; 15: 185–191.
19. Alvaro EM, Siegel JT, Crano WD, et al. A mass mediated intervention on Hispanic live kidney donation. *J Health Commun* 2010; 15: 374–387.
20. Vidal Blandino M, Gentil Govantes MA, Cabello Chaves V, et al. Information channels and the dynamics of uptake of living kidney donors: a retrospective study in a reference area. *Transplant Proc* 2011; 43: 2157–2159.
21. Schweitzer EJ, Yoon S, Hart J, et al. Increased living donor volunteer rates with a formal recipient family education program. *Am J Kidney Dis* 1997; 29: 739–745.
22. Connelly JO, O'Keefe N, Hathaway D, et al. Impact of a human interest video on living-donor kidney donation rates. *J Biocommun* 1999; 26: 7–10.
23. Thornton JD, Alejandro-Rodriguez M, Leon JB, et al. Effect of an iPod video intervention on consent to donate organs: a randomized trial. *Ann Intern Med* 2012; 156: 483–490.
24. Chen HM, Zhong-Kai H and Xiao-Bo L. Effectiveness of YouTube as a source of medical information on heart transplantation. *Interact J Med Res* 2013; 2: e28.
25. Andreassen HK, Bujnowska-Fedak MM, Chronaki CE, et al. European citizens' use of E-health services: a study of seven countries. *BMC Public Health* 2007; 10: 53.
26. Murugiah K, Vallakati A, Rajput K, et al. YouTube as a source of information on cardiopulmonary resuscitation. *Resuscitation* 2011; 82: 332–334.
27. Desai T, Shariff A, Dhingra V, et al. Is content really king? An objective analysis of the public's response to medical videos on YouTube. *PLoS ONE* 2013; 8: e82469.
28. Madathil KC, Rivera-Rodriguez AJ, Greenstein JS, et al. Healthcare information on YouTube: a systematic review. *Health Informatics J* 2015; 21: 173–194.
29. Steinberg PL, Wason S, Stern JM, et al. YouTube as source of prostate cancer information. *Urology* 2010; 75: 619–622.
30. Sood A, Sarangi S, Pandey A, et al. YouTube as a source of information on kidney stone disease. *Urology* 2011; 77: 558–562.
31. Lee JS, Seo HO and Hong TH. YouTube as a source of patient information on gallstone disease. *World J Gastroenterol* 2014; 20: 4066–4070.

Book Review

Health Informatics Journal
2018, Vol. 24(1) 111–112
© The Author(s) 2017
Reprints and permissions:
sagepub.co.uk/journalsPermissions.nav
DOI: 10.1177/1460458217747383
journals.sagepub.com/home/jhi
SAGE

Information Infrastructures within European Health Care: Working with the Installed Base (Health Informatics)

By Margunn Aanestad, Miria Grisot, Ole Hanseth and Polyxeni Vassilakopoulou (eds)
Hardcover: 263 pp.
Publisher: Springer; 1st ed., 2017 edition (22 May 2017)
Language: English
ISBN: 3319510185

Reviewed by: *Gunnar Ellingsen, Telemedicine and E-Health Research Group, UIT – The Arctic University of Norway, Norway*

For decades, the notion of eHealth has carried great promise of improved efficiency and quality of care through information technology (IT)-based capabilities. Along these lines, several strategies have been put into place to implement eHealth in practice. However, a recurring challenge is how eHealth technologies have proven considerably more complex and time-consuming to implement than initially anticipated. This is the starting point for the highly knowledgeable book 'Information Infrastructures within European Health Care: Working with the Installed Base' that deals with the on-going challenges and strategies of implementing eHealth technologies in real clinical practice across Europe.

With a broad approach to the notion of eHealth, the editors include technologies such as Electronic Health Record systems (EHRs), which play a central role in health institutions, picture archiving and communication systems (PACS), radiology information systems (RIS), computerised physician order entry (CPOE), electronic medication management systems (EMMS) and laboratory systems (LAB). A common pattern is that many of these technologies frequently have faced serious problems when confronted with real practice.

The book is organised in three sections: Part I 'Information Infrastructures in Healthcare' presents the empirical domain of the book, the context of eHealth infrastructures, the core theoretical concepts, and the cross-case analysis of the cases. Part II 'E-Prescription Infrastructures' contains six chapters analysing various European experiences with putting in place eHealth infrastructures. The empirical studies on e-prescription are from Spain, Norway, Greece, the United Kingdom and Germany. Part III 'Governmental Patient-Oriented eHealth Infrastructures' contains five empirical chapters on governmental platforms for patient-oriented eHealth services from Spain, Norway, Denmark, Sweden and Italy. The organisation of the book is excellent, as the conceptual backbone of the book on Information Infrastructure is assembled in part I together with the associated cross-case analysis. This increases the readability of the book with a clear focus on the conceptual underpinnings and analysis.

Conceptually, the authors apply an information infrastructure perspective, which has proven very useful for analysing the emergence of large-scale interconnected systems. While such a perspective is well rehearsed in the IS research literature, the authors take a refreshingly deep dive into one of the central elements of an information infrastructure, namely, the installed base. Both physically and conceptually, the installed base represents the existing technologies and practices. The authors argue that we always need to take the installed base into account in all eHealth implementation processes for ensuring a successful organisational outcome. This is very interesting due to the stark contrast of such a strategy with prevalent managerial strategies in many eHealth projects, recognised by high ambitions of sweeping change and replacement of out-of-date technology. The authors elaborate how further evolution of the installed base creates a paradox as it entails building on the installed base and transforming it at the same time – that is, adapting too much to the existing installed base may strengthen its irreversibility and hinder change, while disregarding it may limit the initial utility of any initiative and impede growth. Actually, finding the right balance between stability and change should thus be a key question in eHealth implementation projects.

The empirical cases are carefully selected across European healthcare. They centre on two topics: (a) E-prescription solutions that support the electronic flow of information related to prescribed medications, and (b) the development of patient-oriented eHealth services. The stringent selection of empirical cases creates coherency both among the various cases and within the book as a whole. As such, the empirical cases lay the groundwork for a coherent cross-case analysis that is informative, innovative and convincing. Another intriguing aspect is that the broad selection of empirical cases across Europe demonstrates how qualitative methods – frequently associated with local studies – can be very useful for comparable studies on a large scale as well.

With this book, the editors have done an eminent job of selecting a highly competent author team from different countries across Europe, each highlighting the topic of eHealth from a representative European country. In turn, the authors have exploited the heterogeneous empirical narratives as a foundation for targeted (and homogeneous) conceptual analysis of the installed base concept from the information infrastructure field. Clearly, the book should be mandatory reading for scholars and students in the information systems field as well as for policymakers across Europe.