The Promise of New Technologies in an Age of New Health Challenges
A.J. Maeder et al. (Eds.)

doi:10.3233/978-1-61499-712-2-152

A Population-Level Data Analytics Portal for Self-Administered Lifestyle and Mental Health Screening

Xindi ZHANG [a], Jim WARREN [a,1] Arden CORTER [b] and Felicity GOODYEAR-SMITH [b]
[a] *Department of Computer Science* [b] *Department of General Practice and Primary Health Care, University of Auckland, Auckland 1142, New Zealand*

Abstract. This paper describes development of a prototype data analytics portal for analysis of accumulated screening results from eCHAT (electronic Case-finding and Help Assessment Tool). eCHAT allows individuals to conduct a self-administered lifestyle and mental health screening assessment, with usage to date chiefly in the context of primary care waiting rooms. The intention is for wide roll-out to primary care clinics, including secondary school based clinics, resulting in the accumulation of population-level data. Data from a field trial of eCHAT with sexual health questions tailored to youth were used to support design of a data analytics portal for population-level data. The design process included user personas and scenarios, screen prototyping and a simulator for generating large-scale data sets. The prototype demonstrates the promise of wide-scale self-administered screening data to support a range of users including practice managers, clinical directors and health policy analysts.

Keywords. Data analytics, mental health, population health, primary care, screening

Introduction

Increasing computerisation in healthcare delivery provides a wealth of 'Big Data' for analysis. In the US context, Bates et al. [1] characterise the opportunity primarily in terms of reducing costs, with key use cases for Big Data in high-cost patients, readmissions, triage, patients with worsening conditions, adverse events, and diseases affecting multiple organ systems. In primary care, there is potential for substantial long-term health gains by moderating risky behaviours (such as smoking, alcohol misuse, problem gambling and physical inactivity) and treating mental health issues including anxiety and depression, as well as anger and abuse. To gain these benefits requires detection in the first instance, and longer-term tracking for assessing needs and trends at the population level.

The electronic Case-finding and Help Assessment Tool (eCHAT) has been developed to provide systematic screening for risk behaviours and mental health issues [2]. eCHAT is designed to be operated directly by the health consumer. Screening questions assess potential issues and are followed up by validated assessment tools if

[1] Corresponding Author.

necessary (e.g. the 9-item Patient Health Questionnaire–Depression, PHQ-9 [3]). Further, eCHAT asks a 'help' question in each identified problem area, directly asking the health consumer if they want help with the issue (not at all, at the present time, or later). In the present deployment context, eCHAT is completed prior to visiting a primary care physician (often in the waiting room using a tablet computer provided by the practice) and a summary report is transmitted directly to the doctor's electronic medical record (EMR) system for discussion in the consultation.

The intention is to roll-out eCHAT widely (in its home country of development, New Zealand, in the first instance). This would result in the potential for screening results being available in aggregate at many levels, including the individual practice, as well as to wider networks, regions and nationally. In this paper, we describe the design of a prototype data analytics portal to support use of population-level eCHAT data.

1. Determining Requirements for the Portal

Requirements for the data analytics portal were formulated by analysis of the data collection tool (eCHAT), its data (screening results) and of the intended use of the data.

As mentioned, eCHAT asks the health consumer a series of questions around lifestyle and mental health issues. To keep this computerised interview relevant and as short as practical, the questions on each topic (module) follow a general sequence:

- Screening questions. If these are answered in the negative then the rest of the module is skipped (e.g. if you say you do not drink, then no further questions on the effect of alcohol on your life are asked).
- Discovery questions. Questions where the answers are part of the characterisation of any potential problem in this area.
- Assessment questions. Questions that combine to form a score from a validated instrument (e.g., PHQ-9; Alcohol, Smoking and Substance Involvement Screening Test, ASSIST [4]).
- Help questions. To reduce false-positive assessments and give an indication of patient preference, a module that had a positive screen/assessment finishes by asking if the health consumer wants help with this issue: "no"; "yes, but not today"; or "yes".

There are some overlaps among these types of questions. Taking the depression module as an example, the first two questions of the PHQ-9 serve as the screening questions; and the 9 questions of the PHQ-9 are the totality of the discovery questions (with the last PHQ-9 question, regarding potential self-harm, having specific significance), as well as constituting a set of assessment questions. Furthermore, not all modules have assessment tools, and some screening questions (e.g. regarding concern about sexual orientation in youth) may of themselves constitute the discovery.

At the time of the portal design (2015), eCHAT had recently been expanded with a set of sexual health questions intended for youth. This modification of eCHAT had been field tested at a secondary school in a low-income area of Auckland metro, resulting in 107 interview records collected over four days, from 45 male students and 62 female students (median age 16 years). These records were analysed in concert with

eCHAT technical specifications to develop a full understanding of data domains and dependencies. Additionally, possible research questions specific to these data were discussed between the technical (first two authors) and psychology/clinical (latter two authors) researchers. Note that development, field testing and analysis with respect to expansion of eCHAT for youth sexual health was undertaken in accordance with a protocol approved by New Zealand's 'Northern A' Health and Disability Ethics Committee (NTY/11/10/102).

To further elicit the data analytics requirements and envision the usage of the portal, we followed the interaction design practices of developing personas ("hypothetical archetypes of actual users" [5], p. 123) and scenarios (presentations of "possible ways to use a system to accomplish some desired function" [6], p. 34). We developed a set of personas representative of intended future users of population-level data at various levels of aggregation from an individual practice to nationwide. We then created scenarios for the personas using the portal in ways typical of their roles and responsibilities as they might if eCHAT data were available on a large scale. These personas are summarised in Table 1; Figure 1 shows an example scenario.

Table 1. Data analytics portal user persona summary

Persona Role	Relevant Aspects of Their Job	Example Specific Interest
Practice Manager	Oversees resource allocation in her practice of 3,000 enrolled patients and 3 full-time equivalent primary care physicians	Assessing need for providing specialised sexual health support to patients
Primary Health Organisation (PHO) Clinical Director	Keeping the doctors in his network abreast of evidence based healthcare and clinical variation	Tracking preventable cardiovascular risk factors in Māori patients, including smoking
Ministry of Health Public Health Policy Analyst	Researching and monitoring issues to provide evidence-based policy advice relating to mental health	Following trends in youth depression to address media prompted by a high-profile youth suicide

Review of the personas and scenarios revealed that future users would be interested in far larger data collections than the 107 records that were to hand. For instance, if eCHAT were successfully rolled out to all decile 1 (lowest income) schools in New Zealand and used routinely we would expect approximately 50,000 interview records per year; this would be the data set of interest for our Ministry of Health Policy Analyst persona investigating a question about youth depression. To address this, we opted to create a data simulator. This software generates cases using probability distributions. These were initially programmed with two profiles: one modelled on the distribution in the youth data set; and another with expert (latter two authors) estimates for a general population. The simulator follows the 'rules' of an eCHAT interview in that if a randomly-generated case is negative for the screening questions, then discovery, assessment and help questions for that module are left null.

With the above information to hand, design and implementation of a prototype portal was undertaken (principally by the first author) with periodic feedback from the research team (all authors). Many of the output graphs were initially prototyped in Microsoft Excel prior to being programmed for fully-automated production.

It is nearing the end of the year and Amy is reflecting on the operational aspects of the clinic as she begins to plan for the next year. The three GPs who work at the practice have commented that the number of young people who want help with their sexual health and sexual orientation is increasing, but they lack the expertise and time to offer them long-term support. She wonders if demand is high enough to warrant hiring a sexual health specialist as a part-time or full-time staff member.

The eCHAT tool has been in regular use with her clinic for the last year, so Amy opens the data analytics portal and loads the dataset for her clinic. She decides to focus on youth so she filters by age, selecting 13 as the lower boundary and 18 as the top boundary. Then she navigates to the sexual health domain and begins by looking at how many of the teenagers at the clinic are sexually active. Around 30% admitted to being sexually active and she finds that of those, nearly all admit to feeling at risk of a sexually transmitted disease or pregnancy. "How many patients does that equate to?" she wonders. She changes the labels on the display from percentages to number of patients and discovers that 50 patients have indicated willingness to receive immediate help. She repeats this process to find out the number of teenagers who want help with their sexual orientation concerns or unwanted sex and is surprised by the large number of teenagers struggling with these issues. Amy uses the figures obtained from the data analytics portal to estimate how many hours per week a sexual health specialist/counsellor is required and how much it will cost the practice so she can incorporate it into the clinic's budget.

Figure 1. Data analytics portal usage scenario with Practice Manager persona Amy Simpson.

2. Results

The data analytics portal prototype was built using Shiny, an R package used to build interactive Web applications [7]. A Shiny application is comprised of two components that work in conjunction with each other: a server file that provides instructions for the server, allowing it to create output that responds to user inputs; and a user interface file that acts as an HTML interpreter. The prototype required customised features so additional HTML, CSS and JavaScript code was created to work alongside the R code.

With this architecture, the data analytics capability presents to users as a Web portal operating in their browser, and thus is easily scalable for wide delivery to a range of user types as per our personas. Within the scope of the present study, we did not prototype the security model. However, the personas provide guidance for development of role-based access levels.

The data simulator was developed using Python (Python Software Foundation) with a community-contributed module XlsxWriter [8] to output directly to a Microsoft Excel file format. The examples below use a simulated cohort of 5,000 interviewees generated following frequency distributions derived from the 107 real youth eCHAT field study participants.

Figure 2 shows an example screen from the data analytics prototype. The left-hand side controls data access filtering including the domain of analysis (i.e. which eCHAT module), the range of interview dates, and case demographics (age range, gender and ethnicity). The main section of the screen has a tabbed control. Tabs vary by the domain of interest. For a module with an assessment, such as the depression results shown, a first tab shows frequencies of positive and negative screening. The second tab, as per the figure, shows the frequency distribution for the help question result (using a traffic light colour model) joint with the major ranges of the assessment (level of

depression by PHQ-9 in this case). In this case, a distribution on about 500 cases is shown as the simulator was set to have around a 10% rate of interviewees screening positively and following the branch to complete the full PHQ-9. An additional control on the prototype allows shifting between percentage frequencies and frequency counts. Further tabs give additional data including tabular summaries. As can be seen from this data, youth often did not want help with their depression. Note that while the frequency of asking for help with depression for cases with a positive screen is based on the real data, the correlation between asking for help and severity of depression is *not* modelled in the simulator (see more on this limitation in the Discussion section).

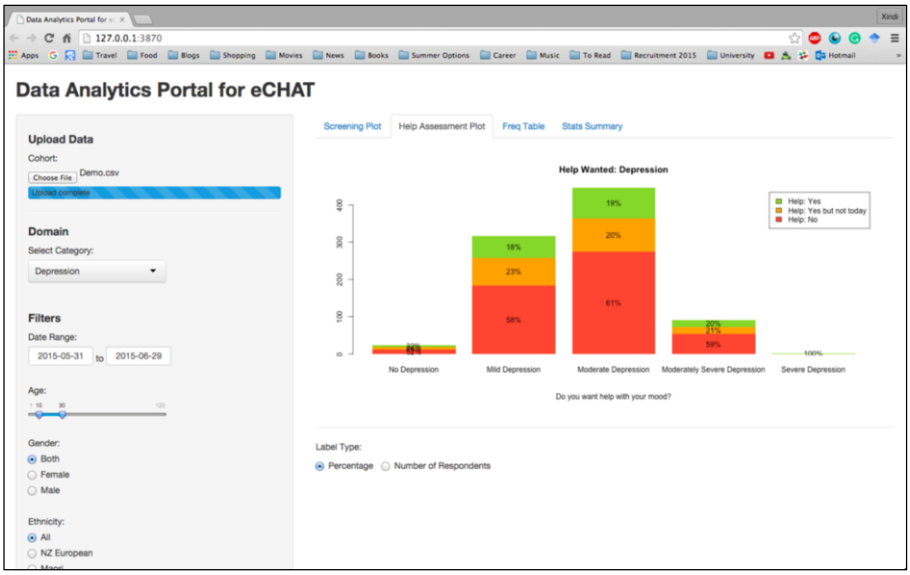

Figure 2. Data analytics prototype screen showing help wanted by depression level

Figure 3 shows a portal screen with the sexual health domain selected. With this domain selected, the left-hand controls adapt to allow selection of sub-domains (sexual orientation, sexual activity or unwanted sexual contact). The main portal provides options within the tabs appropriate to the selected subdomain. The results shown indicate a surprisingly low number of respondents indicating that they are sexually active. As this frequency was derived from the real data, the result served to reveal a problem in the wording of the screening question. The sexual activity screening question explicitly mentions sex involving “your mouth, vagina or bottom” but it does not mention “penis” and thus may have been misinterpreted by many male respondents. Although the wording of the question had been vetted and found clear by a mixed-gender youth focus group, it was flawed in the actual flow of the questionnaire. This shows the value of data analysis to the ongoing refinement of any data collection instrument.

Lifestyle and mental health problems are known to sometimes be correlated [9] or to occur in related clusters [10]. A further area to extend the data analytics platform would be toward analysis of associations between (and among) eCHAT assessments per individual. Some basic design concepts for this, such as use of scatterplot, were prototyped with Excel but not implemented into the R/Shiny prototype.

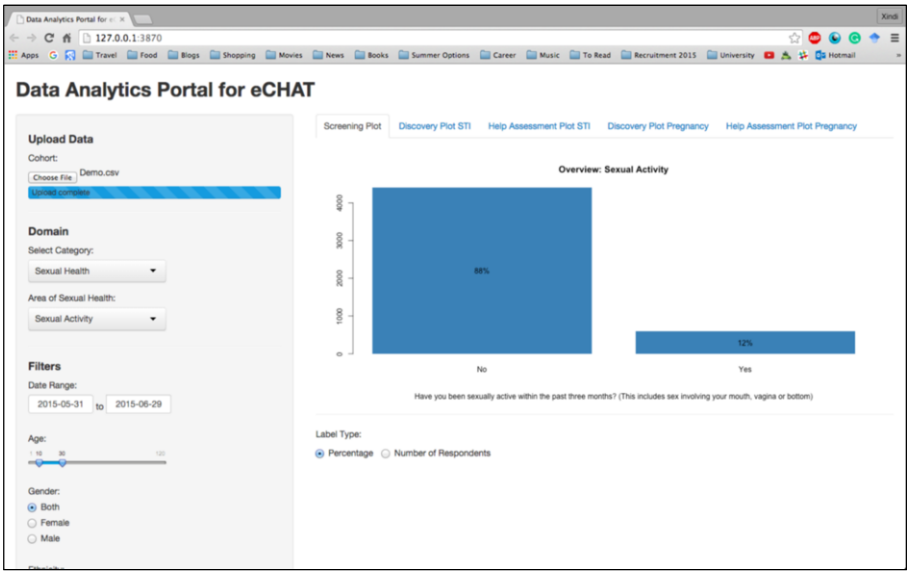

Figure 3. Prototype screen showing screening results for sexual activity.

3. Discussion

This paper reports on the development of a prototype data analytics portal to support use of population-level screening data as applied to eCHAT. This systematic screening tool is a self-administered computer-based instrument measuring lifestyle and mental health issues and has potential for wide-scale deployment in primary care, including settings tailored for youth (e.g., secondary schools). For youth cohorts, a module of sexual health questions was developed and field tested. The data analytics portal prototype demonstrates easy-to-use data filtering and summary displays that would suit a range of future user personas envisioned as having an interest in the aggregate screening results of eCHAT.

There are specific architectures for management of large-scale health data that are relevant to consider as the data analytics capability moves from prototype to production use. Notably, the 'integrated biology and the bedside' (i2b2) [11] is proven as a software framework to create 'data marts' for analysis and could map to the eCHAT case. Moreover, we should not limit our vision to eCHAT used in isolation, given the excellent potential for extensive data linkage through the Statistics New Zealand Integrated Data Infrastructure (IDI) [12] for combination with social determinants of health, health outcomes and other variables. Nonetheless, we feel that the prototyping exercise presented herein is useful for understanding minimal requirements for a system built on these more extended infrastructures. Moreover, it appears there are many useful questions held by typical users that could be answered with relatively simple capabilities, particularly if they are available with the ease-of-use of a customised tool such as we have demonstrated with the present prototype design.

One area in need of considerable functional expansion is the ability to examine associations among screening variables. In visual terms, simple scatterplots of two variables are a starting point. To address multivariate associations, and to cope with the

larger end of our anticipated scale of data (e.g. for national level users), visual representations such as data cubes with interactive pan-and-zoom manipulation of the range of display [13, 14] are promising and increasingly mature. It is unclear whether ease of use can be maintained while providing support for the range of questions (and related visualisation and analysis needs) likely to arise from users of the data collection. At some point an 'export' to a traditional statistical package would be needed to address more advanced analytical goals; however, some easily accessed in-built functionality for analysis of associations seems warranted to address cases in keeping with our personas and scenarios. Further, some of the easy-to-use filtering features may be a convenient starting point for selecting a data subset for export to another package.

Related to the display of association in the data, further research is needed to create a data simulator that can generate test data with the expected associations among variables. The present simulator generates values independently. Therefore as an example, a problem gambler is no more likely to be depressed than a non-gambler, which is contrary to existing findings [10]. Modelling this will require a more sophisticated case generation module with an explicit representation of problem correlations and clusters.

There are moves to employ eCHAT in a number of different vulnerable populations including youth in secondary school and primary care settings, and ethnic minorities in New Zealand such as Chinese and Korean. It is also planned to adapt the tool for non-clinical settings, especially for young people, who can self-administer the test and receive tailored self-management interventions including psychoeducation and links to e-therapy. Analyses of population-level data can assist in estimating the prevalence of mental health and lifestyle issues, gain an understanding of their inter-relatedness and assess unmet service need.

Acknowledgments

This project was assisted by funding from the Maurice & Phyllis Paykel Trust (grant project number 370679). We would like to thank the clinical staff and students of the secondary school for their active participation in this project.

References

[1] Bates, D.W., Saria, S., Ohno-Machado, L., Shah, A. & Escobar. G. Big data in health care: using analytics to identify and manage high-risk and high-cost patients. *Health Affairs* **33** (2014), 1123-1131.

[2] Goodyear-Smith, F., Warren, J., Bojic, M. & Chong, A. eCHAT for lifestyle and mental health screening in primary care. *Ann Fam Med* **11** (2013), 460-466.

[3] Kroenke, K., Spitzer, R.L. & Williams. J.B. The PHQ-9: validity of a brief depression severity measure. *J Gen Int Med* **16** (2001), 606-613.

[4] Humeniuk, R., Ali, R., Babor, T.F., Farrell, M., Formigoni, M.L., Jittiwutikarn, J., de Lacerda, R.B., Ling, W., Marsden, J., Monteiro, M., Nhiwatiwa, S., Pal, H., Poznyak, V. & Simon, S. Validation of the Alcohol, Smoking And Substance Involvement Screening Test (ASSIST). *Addiction* **103** (2008), 1039-1047.

[5] Cooper, A. *The Inmates are Running the Asylum: Why high-tech products drive us crazy and how to restore the sanity.* Indianapolis: Sams (1999).

[6] Weidenhaupt, K., Pohl, K., Jarke, M. & Haumer, P. Scenarios in system development: current practice. *IEEE Software* **15** (1998), 34-45.

[7] RStudio. Teach yourself Shiny. Retrieved from http://shiny.rstudio.com/tutorial/ (last accessed 10 June, 2016).

[8] McNamara, J. Introduction to XlsxWriter. Retrieved from http://xlsxwriter.readthedocs.org/introduction.html (last accessed 10 June, 2016).
[9] Saluja, G., Iachan, R., Scheidt, P.C., Overpeck, M.D., Sun, W. & Giedd, J.N. Prevalence of and risk factors for depressive symptoms among young adolescents. *Arch Pediatr Adolesc Med* **158** (2004), 760-765.
[10] Goodyear-Smith, F., Arroll, B., Kerse, N., Sullivan, S., Coupe, N., Tse, S., Shepherd, R., Rossen, F. & Perese, L. Primary care patients reporting concerns about their gambling frequently have other co-occurring lifestyle and mental health issues. *BMC Fam Pract* **7** (2006), 25.
[11] Murphy, S.N., Weber, G., Mendis, M., Gainer, V., Chueh, H.C., Churchill, S. & Kohane, I. Serving the enterprise and beyond with informatics for integrating biology and the bedside (i2b2). *J Am Med Inform Assoc* **17** (2010), 124-130.
[12] Statistics New Zealand. *An overview of progress on the potential use of administrative data for census information in New Zealand: Census Transformation programme*. Available from www.stats.govt.nz (2014).
[13] Stolte, C., Tang, D. & Hanrahan, P. Multiscale visualization using data cubes. *IEEE Transactions on Visualization and Computer Graphics* **9** (2003), 176-187.
[14] Liu, Z., Jiang, B. & Heer, J. *imMens*: Real-time visual querying of Big Data. *Computer Graphics Forum* **32** (2013), 421-430.