

The Challenge of Evaluating Electronic Decision Support in the Community

Jim Warren^{1,2}, Rekha Gaikwad², Thusitha Mabotuwana¹, Mehnaz Adnan¹, Tim Kenealy³,
Beryl Plimmer¹, Susan Wells², Paul Roseman⁴ and Karl Cole⁴

1 Department of Computer Science

2 Section for Epidemiology and Biostatistics

3 Department of Medicine

The University of Auckland

Private Bag 92019, Auckland 1142, New Zealand

jim@cs.auckland.ac.nz

4 ProCare

PO Box 105 346, Auckland, New Zealand

Abstract

There is clear room for improvement in all existing clinical decision support systems (CDSS) to more closely meet the needs of their target users and work processes. Ongoing evaluation serves not just to document benefits, but also to provide feedback to the CDSS engineering process to create more usable and effective tools. This paper discusses the challenges of achieving evaluation data that provides detail on system use at the point of care with a focus on improvement of decision support in a community setting such as a General Practice clinic. After describing some results of CDSS evaluation from the literature, as well as our own recent efforts at laboratory evaluation of the PREDICT CVD/Diabetes tool, we identify specific challenges that arise in CDSS evaluation. We suggest and discuss several approaches that may lead to facilitated and improved data gathering to help CDSS developers understand how their tools are being used and hence how to improve them, including: (a) the challenges and opportunities afforded by automated logging of CDSS usage; and (b) the value of an electronic health record interchange standard for the creation of realistic test cases.

1. Introduction

Clinical decision support systems (CDSS) are considered the pinnacle for achieving Evidence Based Medicine [1]. How do we reliably create effective clinical decision support systems (CDSS)? How do we make CDSS that are used and valued because they make doing the 'right things' easy, that are not burdensome to healthcare delivery, and that, most importantly, promote a high quality of care in accordance with evidence based clinical practice guidelines?

Kawamoto et al [2] screened over 10,000 articles to select 88 papers on 70 studies that assessed the ability of decision support systems to improve clinical practice. They identified four features as independent predictors of improved clinical practice: automatic provision of decision support as part of clinician workflow; provision of recommendations rather than just assessments; provision of decision support at the time and location of decision making; and computer based decision support. These findings provide a valuable touchstone for achieving CDSS success; however, they are merely significant correlates to success. They are not necessary and sufficient features; and they provide minimal direct insight on how well the results may transfer to another context, or about the maintainability and adaptability needed for sustainable solutions. Absence versus presence of all four of the independent predictors moved success rates from an observed mean of 46% to 94%, but we must consider that publication bias may be masking a larger underlying failure rate and/or may be biasing what sort of implementations get reported.

Chaudhry et al. [3] performed a systematic literature review on evidence of improved healthcare from health IT generally. From 257 included articles they found that approximately 25% of the studies were from just four benchmark institutions that implemented internally developed systems; and only nine studies evaluated multifunctional, commercially developed systems. This concentration of results in a few institutions that have undertaken exceptional, decades long, programmes of internal development led them to question whether widespread achievement of similar health IT benefits can be expected at all.

The lack of uptake of CDSS has been recognised by the American Medical Informatics Association (AMIA), which in June 2006 released *A Roadmap for National Action on Clinical Decision Support* [4], acknowledging the “limited extent” to which clinical decision support is being leveraged in the US. This roadmap puts forth three ‘pillars’ for realising the potential of CDSS for more widespread positive impact on health outcomes: best knowledge available when needed; high adoption and effective use; and continuous improvement of knowledge and clinical decision support methods. It has been suggested that the first of these pillars (i.e., availability) can be addressed through adoption of interoperability standards as per the Healthcare Services Specification Project (HSSP) and with associated use of Service Oriented Architecture (SOA) [5]. One may, or may not, be optimistic about such a technological ‘magic bullet’ for CDSS availability; addressing the latter two pillars, however, will require extensive and detailed evaluation of how CDSS are being used, with particular attention to barriers to use that are encountered.

This paper identifies challenges to successful evaluation of CDSS at the point of care with a focus on decision support aimed at improving chronic disease management in the community. We discuss several approaches that may lead to facilitated and improved gathering of evaluation data to help CDSS developers understand how their tools are being used and hence how to improve them.

2. CDSS Evaluation

It is instructive to consider a cluster-randomised controlled trial of the effect of a CDSS for management of asthma and angina in adults, based on 60 General Practices in northeast England, reported in 2002 [6]. The trial in fact found no improvement in process of care or health outcome, and indicated that system usage levels were low. A qualitative interview study associated with the trial [7] revealed that potential users (GPs and nurses) perceived several barriers to uptake of the system: timing of the guideline trigger, ease of use of the system, and helpfulness of the content. The issues around the evaluation attracted extensive, and sometimes conflicting, discussion (best viewed through the ‘Rapid Responses’ section of the *BMJ* on the Web). The outcomes highlighted the importance of timely training and, in general, of achieving a good fit of the CDSS user interface to the way the clinical user is trying to use it. The case serves to demonstrate that an extensive, well-resourced CDSS deployment can manage to fail quite comprehensively in reaching the requirements for enhancing chronic disease management in General Practice.

In welcome contrast to that rather gloomy story from the UK, the uptake of the PREDICT CVD/Diabetes risk assessment and management support tool (hereafter ‘PREDICT’) has been a significant success. As of January 2008, around 1,500 doctors and nurses are currently using PREDICT with approximately 3,000 risk assessments being conducted per month. The participating primary health care organisations that have provided written consent for the use of anonymised patient CVD risk data for CVD risk prediction research have contributed over 100,000 risk assessments conducted on over 50,000 patients to a national CVD research cohort.

PREDICT fits well with the success factors identified by Kawamoto et al. [2] with respect to providing computer-based automated risk assessments and associated specific recommendation. On the other hand, PREDICT is at some variance from Kawamoto et al’s ideal in that it requires a manual invocation (albeit one that is available through the GP’s PMS) and that a PREDICT session often requires additional data entry. This is due to two key issues: the challenge of extracting narrative free text that has the same semantic validity as an evidence-based guideline variable; and it highlights an evidence-practice gap wherein the variables necessary for CVD risk assessment and management are not being collected in routine care. Thus additional time is required to conduct this preventive care process.

PREDICT is sufficiently widely available in New Zealand that, while the usage figures are substantial, it is equally clear that clinicians could elect to use the software much more than they do. Moreover, among those using PREDICT, the extent to which they are using it as the designers intended within patient consultations is not entirely known. This motivates more in-depth study of how the CDSS is being used at the point of care.

We devised a set of protocols, “How general practitioners perceive and use computerized risk-based chronic disease management advice” and “Use of PREDICT interface and functionalities during the patient-provider consultation,” approved by the University of Auckland Human Participants Ethics Committee. These protocols were designed to bring the in-depth insights possible with video recording of PREDICT sessions to bear on the issue of just how GPs and nurses use the software when in consultation with a patient. With video, we would not need to rely on human memory to know, for instance, which recommendations were discussed with patients, and in what terms. Moreover, combining video of the clinician and patient with a full-motion log of screen activity, we can reconstruct the exact allocation of time within a consultation amongst speaking with the patient, review of recommendations, data entry, system latency and error recovery.

In terms of usability testing platform, we employed Morea usability testing software by TechSmith, along with a system of one computer mounted camera (“Web cam”) and a second camera positioned to capture the interaction from the patient perspective. With Morea providing an underlying recording function, we ran Medtech32 and from there launched PREDICT as an ‘Advanced Web Form.’

We opted to use medical actors (actors experienced with role-playing for medical certification exams) in lieu of patients and to conduct the work in a ‘laboratory’ setting using consulting rooms at The University of Auckland’s School of Population Health. This was done to remove the burden of achieving patient consent and to allow a concentrated session where all presenting cases were targeted to specific aspects of PREDICT’s recommendation capability. Moreover, the use of a controlled setting facilitated the installation and configuration of the usability software and cameras. In a first round of study, six GPs were recruited, with three of these GPs being individuals associated with the development and evaluation of PREDICT (including two of the authors) who were taken to reflect ‘model’ usage of the system. A second round of investigation was started (but not completed) involving recruitment of GPs and nurses from the PREDICT user population. Figure 1 shows a screen display from the usability testing software reviewing the log from a consultation.

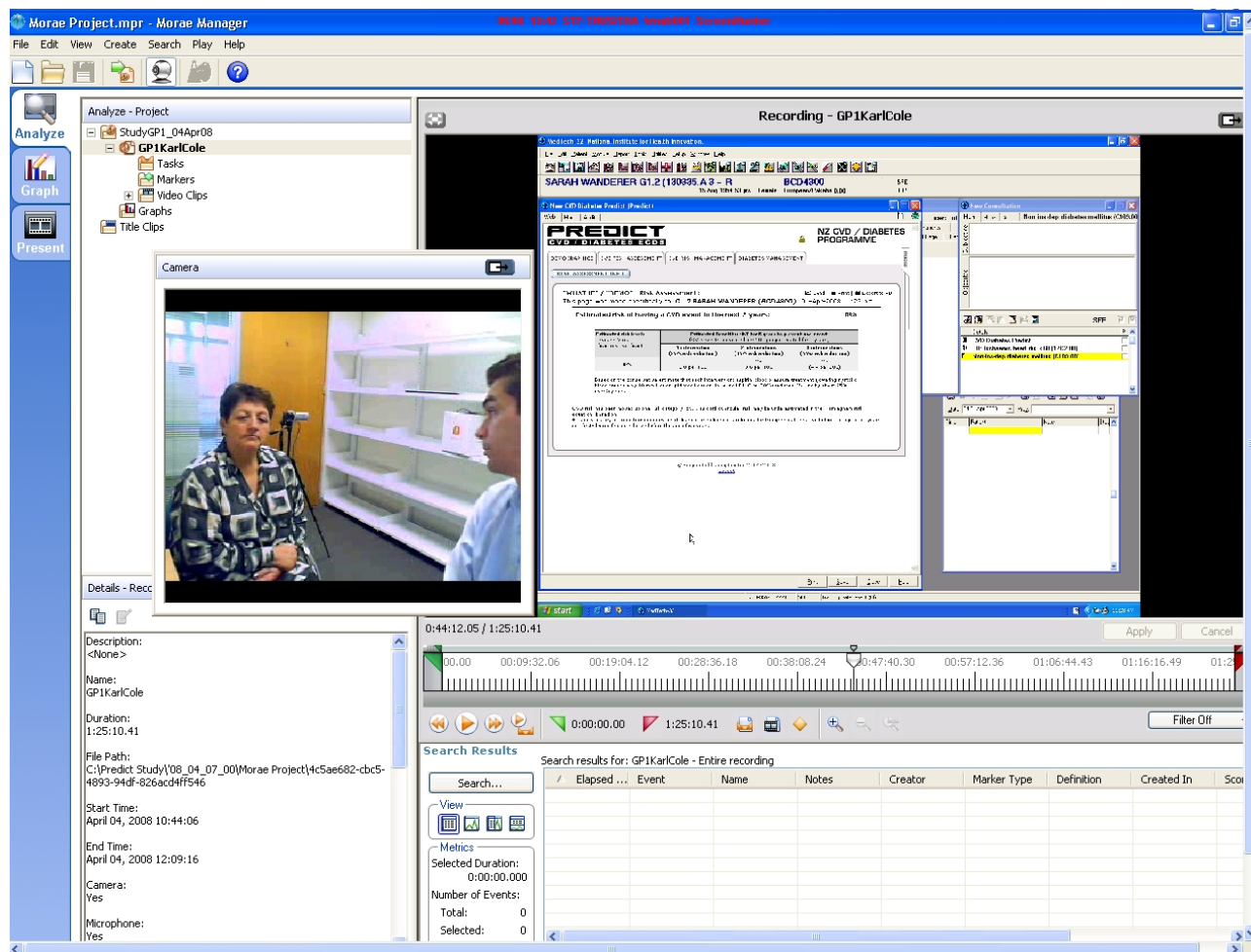


Figure 1 – Morea screen capture with inset video of user; GP (Karl Cole) using PREDICT with a medical actor.

The usability experiment was designed around three cases, each portrayed by a different actor, and timed to fit within one hour of experimentation (15 minutes per case and some overheads in explaining the scenarios and debriefing). The cases were diverse in ethnicity, gender, level of CVD risk and degree to which the patient was controlling that risk. Each case was ‘scripted’ with both a conventional story (the path the patient had taken in getting to this point) and an associated medical record (e.g., current medications, lipids, etc.). In the first round of investigation a relatively minimal record was present in the Medtech32 database and other data was entered by the GP as part of the video-recorded study. For the second round of investigation the clinical profiles of the three cases were refined, as was each story, and more comprehensive data was entered into Medtech32. The second round was aimed at achieving a tighter focus on how clinicians utilised the patient education and clinical recommendation/action functions from PREDICT.

The experimentation was halted early in the second round of investigations due to a number of issues. These issues inform our discussion of challenges to CDSS evaluation in the next section.

3. Challenges to Evaluation

3.1. Rationalisation and Recollection

A compellingly practical way to gather data on CDSS usability is simply to ask. This can take the form of questionnaires (online or paper), interviews or focus groups. Response rate and recruitment are perennial problems here, but possibly not as acute as with observational studies (see 3.2 below). This approach is good in that it overcomes some of the intrinsic problems with evaluation of CDSS in the community. Not every patient presenting to a practice is a candidate for the use of a particular CDSS; depending on the domain of the system, and the nature of the case load at the practice, it may be used just a few times per week. But a questionnaire, interview or focus group session can be scheduled at the end of an extended period of use whereupon the clinician users have had time to form an opinion of the system's performance.

Limitations of these methods lie in the lack of objective data. The questions that guide these methods are subject to validity concerns – are the questions understood the same way as intended? – do the questions bias the answers (i.e., are they 'leading questions')? More perniciously, subjects may not be able to (or may choose not to) separate their agendas from their responses. The line of reasoning can be quite benign, such as: "This system espouses evidence based medicine (EBM); EBM saves lives; using the system saves lives; the system is worth using; I felt it was worth using; I'll respond favourably to questions about its usability." Or, conversely, the subject may be negatively predisposed to the technology for any of a spectrum of reasons. Asking about overall satisfaction or desire to use the system in the future, while valuable for predicting sustainable use of the system, is inviting this sort of rationalised assessment. In this regard, focus groups, through the dialogue amongst peers with diverse views, may have the best potential (as compared to questionnaires or single-subject interviews) to 'unpack' the factors that lead to a negative assessment and reveal system characteristics that can be improved by re-engineering of the product. Also, methods such as *cognitive interviewing* (getting subjects to talk aloud under observation as they provide responses to questions) can be used to test and improve questionnaire and interview questions [8].

Questionnaires, interviews and focus groups are undeniably a mainstay for gathering evaluation data. The power of these methods to give a clear picture of events, however, is greatly amplified when combined with observation of system use [9, pg 348]; that is, it is more useful to the CDSS engineer to see a recording of the event that led to, for example, dissatisfaction with the system, than to just hear a description weeks after the fact. Moreover, direct observation provides quantification of outcomes – such as length of session, frequency of errors – which may be conveyed in broad terms in an interview, but lack the backing of hard numbers.

These methods are still valuable enough to undertake, but it is due to their limitations in informing CDSS improvement that we pursued more direct observation of system use.

3.2. Consent, Recruitment and the Problem with Video

Observational study provides a detail and objectivity that is lost when simply asking a user to recall and describe the experience of using the system; but also observation has many intrinsic flaws. There are many well-known biases that can be induced when subjects are aware that they are observed [10]. In the context of CDSS evaluation, it is essentially impossible to blind the participant to the fact that they are receiving electronic decision support, so the clinician's attitude toward CDSS may influence their degree of effort taken to smooth over, or get hung up on, usability problems.

In the chronic disease management context we encounter additional challenges to observational study in that only a percentage of the general practice caseload will be candidates for a particular CDSS intervention. This adds complexity, error, and significant research staff time costs to efforts to recruit patients in the waiting room, and was part of our motivation for choosing to pursue a laboratory protocol with medical actors. The use of a laboratory context, however, brought its own recruitment problems in that it was difficult to find GPs that wished to commit to investing time in such a protocol (and, as per 3.4 below, this is an expensive option, too). This recruitment bias was sufficiently strong as to make results questionable (i.e., with the prevalence of willingness to participate so low in the GP population, it seems likely that those recruited may have an extreme view, most likely as a CDSS enthusiast).

As a further complication, a one hour laboratory session was far too short a period for the GP to 'forget' about the camera, with an unknown effect on GP behaviour. The best solution to this would be to return to a field setting and keep the cameras rolling for a sufficiently large number of CDSS sessions until the user could forget about the recording; but this opens a host of further issues.

Oakley et al. [11] report the routine video-recording of the efforts of the trauma team over an 18-month period for the resuscitation of children presenting to the emergency department of the Royal Children's Hospital in Melbourne. The method has highly effective in identifying variance to their chosen guideline (with a mean of 5.9 errors per incident). The recording was not, however, without problems – out of 150 trauma resuscitations, 45 were not recorded (either due to technical problems or the staff not turning on the video) and a further 15 were excluded because interference made the video unviewable to the standards of the analysis protocol. In this case, the institutional ethics committee deemed

that patient or parental consent was not required in that the work constituted routine audit; all affected staff agreed to the video recording.

Is such a method transferable to General Practice? Would patients be comfortable with the idea of their sessions with the GP and practice nurse being videotaped as a matter of routine audit? Even if so, Oakley et al point out the effort/cost of the analysis for their 90 video episodes. Furthermore, there are differences between a large hospital setting and the community. In our study, the use of actors obviated patient confidentiality concerns; but a community-based field study is likely to involve multiple sites and the subsequent transport of data to an external facility. In the case of our team, the investigators include both clinical and non-clinical (i.e., IT-oriented) staff and students. Such a situation, which seems the most practical scenario for evaluation of CDSS use in the community, significantly extends the confidentiality risk and exposure of the resultant videos as compared to the case of a hospital doing internal audit.

3.3. Realistic Test Cases and Software Environment

Implementing a realistic test case for a computerised General Practice holds several layers of challenge. The starting point is to achieve coverage of the desired test domain. Between our first and second protocols, we tightened this scope. For instance, in the first round we had a case with a pre-existing heart condition, which was subsequently deemed to be too obvious a case to test the most relevant features of PREDICT. With the general scope agreed, we needed to formulate a completely feasible set of clinical observations for the patient (this could be aided by working backward from real cases). Beyond this, however, each hypothetical patient had to be at realistic point in their history and management. At this point subtle challenges emerge when simulating a healthcare system with the degree of computerisation found in New Zealand.

Sometimes we may wish to simulate a patient that has just 'dropped in' for a first-time visit and has no presently-retrievable details in the PMS beyond what the practice nurse and/or GP have recorded in the past few minutes, or what may be on a piece of paper the patient brought with them. This is a *possible* scenario for PREDICT use, although not an ideal one, and certainly not the *only* scenario we would wish to model. We must also be able to simulate a patient that is making a return visit.

In our second round of usability testing, we simulated an overweight Indian man in his 40's that had presented the previous week with flu, and now was visiting, much recovered, to get a medical certificate for their employer. At this point the GP turns the consultation toward his CVD risk and invokes PREDICT. On our first video-recorded session with the participating GP and medical actor, the GP is stymied that the visit from last week isn't in the Medtech32 system. The visit for flu isn't clinically important for a CVD risk assessment, but it just looks 'wrong' that it's not there. Herein lies a significant challenge for a laboratory study. Can we, shortly before a laboratory session, enter a complete set of clinical transactions on a patient (prescriptions, diagnoses, notes, lab tests and observations), each appropriately back-dated to its simulated time of occurrence? Probably yes, but it is a laborious manual task; and simulating results that came in from external sources may require placing simulated Health Level 7 messages on the network, or forging data directly into the underlying database of the PMS. Moreover, all of this data will be entered with specific real dates. Within a few days, "last week" has moved on, and old data must be expunged and replaced (or a new case created).

3.4. Accepting the Cost

The cost of using health professionals as study participants is high to the point that: (a) it strains research budgets and (b) it generates organisational resistance to funding such costs. Consider GPs (arguably the most relevant subjects for CDSS evaluation studies): the individuals are reasonably well paid, but furthermore tend to be owners of their own businesses or operating as independent contractors for some or all of their professional hours. As such, it is nearly meaningless to suggest that the study be done 'out of hours' (they could have been working at any hour of the week) and the cost needs to be multiplied by around a factor of two or more to accommodate fixed business operation expenses. The resulting hourly rates that must be paid by the evaluation project are high.

The first author has led a number of research evaluations involving direct data collection from GPs. Noting in particular the three studies reported in [12-14], in each of these cases, as well as the present study: (a) the funding body challenged the funding required for GP participation in data collection (or simply dropped it, requiring scale-back and re-allocation of funds); and (b) the amounts were challenged by the relevant institutional review committees, delaying the research and adding to the administrative burden of its conduct. If a study has a very high degree of GP leadership and interest, as with [15], then the cost per se may be better accepted by funders and review boards, but it is still a substantial issue. Moreover, the cost may be somewhat hidden by GPs receiving supplemental funding to participate through organisations that partner in the research (such as Ministry of Health, District Health Boards and Primary Health Organisations [PHOs] in the New Zealand setting). This does not make the cost go away, and perhaps is unhelpful in developing the culture of accepting the full cost of such evaluation.

The cost of clinician time for evaluation activities is an issue for focus groups and interviews, but is most acute for laboratory studies. This cost is lessened in field studies, since the clinicians continue to deliver care while under

observation; but there are likely to be logistic overheads (e.g., if patient informed consent is taken) and time explicitly set aside for debriefing interviews or other measurements. Moreover, when using video, the cost of research staff time also mounts up rapidly; Berry and Hart [16] say it can take up to ten hours to code and analyse one hour of videotape, which is in agreement with the investigators' experience.

4. Achieving the Best Evaluation Data

Berry and Hart [16] review the well-known methods for evaluation of any "expert systems" technology (clinical or otherwise) and identify the following categories, a number of which we have already mentioned: interviews; questionnaires; system walk-through; formal observation; user diaries; system logging; and simple experiments. This taxonomy informs our suggestions below for improving

4.1. Exit Interviews and Focus Groups

Berry and Hart suggest 'user diaries' as a source of evaluation data that is sustainable over long field trial periods – however, user diaries suffer from the problem of users failing to make entries, due to other demands of work as well as simply forgetting. This deficit may be overcome simply by increasing the scale of the data collection (as long as there is no systematic bias in when users do and do not log to their diaries). The *exit interview* offers some of the features of a user diary, and greatly reduces the limitations of a traditional interview that attempts to collect a summary of experience over a long period of time. A brief exit interview conducted immediately after each CDSS session can allow observations on system performance to be recalled and recorded while still fresh, and these details can form the basis for subsequent discussion, such as in a focus group setting. An exit interview strategy will be intensive on research time in inverse proportion to the prevalence of CDSS candidate patients in the case load (i.e., low frequency of CDSS use equates to a lot of time sitting around waiting to do an exit interview). We believe this could form an excellent complement to automated logging (see 4.3) to allow the quantitative measurement of data entry effort and time to be aligned with user impressions on a session by session basis.

4.2. 'Discount' Usability Testing

In keeping with what Berry and Hart call 'system walk-through,' there are accepted benefits to the systematic assessment of the CDSS user interface by usability experts (this is dubbed a 'discount' technique on the basis that the cost is lower than running a formal observational study). While it may be difficult for experts to place themselves in the shoes of novice users, it is often possible for a small number of reviewers to identify the majority of usability flaws in a user interface [9]. For instance, in a *cognitive walkthrough*, a review panel moves through the application of a system for a defined task, and at each and every step of user action the panel answers a series of questions about the suitability of the state of the display to support the user undertaking (being able to see how to undertake, and getting feedback on) the required action step [17]. It is only a supposition, but it seems that an interdisciplinary panel may be needed to support a cognitive walkthrough of a CDSS. Alternatively, we can record 'ordinary' users (e.g., GPs and nurses) performing set tasks on the CDSS while asked to 'talk aloud' (known as *protocol analysis* [9, pg 344]) – and expert analysis is then made from the videorecording.

4.3. Automated Logging from the Vendor/Host

Automated logging (or 'system logging') involves the system itself automatically recording user actions. This can be imposed as a separate software layer by installing and activating a tool such as Morea. However, this entails a non-trivial system overhead, which may be observed in performance degradation, and produces enormous system logs (e.g., perhaps a few Gigabytes per hour); and the results must still be analysed and coded by a human. Automatic logging can be much more efficient and targeted if it is built into the application itself. This, of course, implies collaboration of the vendor in the evaluation process.

A number of important clinical applications in New Zealand and elsewhere are Web based. This places limitations on the granularity of automatic logging that is readily available to the host server (e.g., only being aware of accesses to the server; i.e., when screens are submitted and new screens requested). However, with coding effort from the vendors, it would be possible to create fine-grained logs of down to the level of character entry and including all field navigation, through use of client-side logic embedded in the Web content. This would allow the collection of data to measure, for instance, whether users generally advance through a data entry task top to bottom, or perhaps often hop around (with the latter suggesting a need to re-order the form).

4.4. CDSS-prompted Enrolment and Video-recording

A partial solution to the issues in video-recording of sessions, would be to have the CDSS itself identify a case as a possible candidate for inclusion in usability evaluation and then prompt the clinician user to gain patient consent and (ideally, automatically upon consent) to activate recording. Such a method could also implement sampling strategies and spread the burden of obtaining consents across a wide user base with minimal researcher effort. One downside of

this would be the need to pre-deploy a large amount of audio/video equipment; however, equipment adequate to the task is becoming increasingly ubiquitous for use of applications such as Skype and Webex.

4.5. Exploiting Electronic Data Interchange Standards

The most obvious motivation for electronic interchange of an extended electronic health record (as compared to just a summary) would be to allow a patient to transfer their care from one provider to another as, perhaps, in formally changing PHOs. At present, there is no easy way to do this sort of interchange, and it would seem that the only option for transfer between two electronic practices is to print everything out and then type it into the new system.

Methods such as HL7 Clinical Document Architecture [18] and OpenEHR [19] provide frameworks for defining, as XML documents, sufficiently comprehensive content to implement a record transfer between General Practice providers. These methods would also be exceedingly useful for creating simulated patient data as needed for CDSS evaluation in a laboratory setting. As discussed earlier, at present the task requires extensive manual data entry through the PMS user interface. A method based on electronic transfer would be faster and more maintainable, and would easily allow the data to be back-dated to align with a specific usability testing date.

5. Conclusion

Review of records and conventional study designs allow the assessment of CDSS usage levels and measurement of impacts on both process and, with sufficient scale and duration of study, of patient outcomes. Questionnaires and interviews can also gather aggregate user opinion of systems. These methods, however, lack the detail to provide direction for engineering of more effective and usable decision support tools.

We have identified a number of challenges in CDSS evaluation as it relates gathering data for system improvement using a community based CDSS system. These challenges include the logistical difficulties of creating video-recordings and the high cost of the associated research, both in researcher and clinician time.

We have suggested a variety of methods for improving the collection of data for CDSS improvement, including the use of exit interviews, and modification of the CDSS itself to aid study enrolment and to provide automated logging targeted to usability evaluation hypotheses.

The area of CDSS usability remains under-studied, even as the amount of data on CDSS outcomes is growing. This is probably due to the very challenges that we have highlighted; it may also be due to a bias against publication of qualitative results, which often have the feel of merely expensively collected anecdotes. CDSS success, however, remains an isolated phenomenon when considered across the spectrum of healthcare activity, or even computer-supported healthcare activity. This places a continuing burden on CDSS developers and researchers to gather detailed data on system use and usability.

6. Acknowledgments

This work is supported in part by a University of Auckland Staff Research Fund grant and Tertiary Education Commission Partnerships for Excellence funding. We acknowledge the helpful input of Stewart Wells, Tania Riddell, Chris Wiltshire and Mike Stanbridge.

7. References

- [1] Haynes RB. Of studies, syntheses, synopses, summaries, and systems: the "5S" evolution of information services for evidence-based healthcare decisions. *Evidence Based Medicine* 2006;11:162-4.
- [2] Kawamoto K, Houlihan C, Balas E, Lobach D. Improving clinical practice using clinical decision support systems: a systematic review of trials to identify features critical to success. *BMJ* 2005; 330(7494), 765.
- [3] Chaudhry B, Wang J, Wu S, Maglione M, Mojica W, Roth E, Morton S, Shekelle P. Systematic review: impact of health information technology on quality, efficiency, and cost of medical care. *Ann Intern Med* 2006; 144(10): E12-22.
- [4] Osheroff JA, Teich JM, Middleton B, Steen EB, Wright A, Detmer DE. A roadmap for national action on clinical decision support. *J Am Med Inform Assoc* 2007;14:141-5.
- [5] Kawamoto K, Lobach D. Proposal for fulfilling strategic objectives of the U.S. roadmap for national action on decision support through a service-oriented architecture leveraging HL7 services. *J Am Med Inform Assoc* 2007; 14:146-55.

- [6] Eccles M, McColl E, Steen N, Rousseau N, Grimshaw J, Parkin D, Purves I. Effect of computerised evidence based guidelines of asthma and angina in adults in primary care: cluster randomised controlled trial. *BMJ* 2002; 325(7370), 941-8.
- [7] Rousseau N, McColl E, Newton J, Grimshaw J, Eccles M. Practice based, longitudinal, qualitative interview study of computerised evidence based guidelines in primary care. *BMJ* 2003; 326(7384): 314-21.
- [8] Willis G. Cognitive interviewing revisited: a useful technique, in theory? In (Presser S, Rothgeb J, Couper M, Lessler J, Martin E, Martin J, Signer E, eds.) *Methods for Testing and Evaluating Survey Questionnaires*. John Wiley & Sons, 2004, pp. 23-7.
- [9] Dix A, Finlay J, Abowd G, Beale R. *Human-Computer Interaction: 3rd ed*. Pearson Education Limited, Harlow, England, 2004.
- [10] Draper SW. The Hawthorne, Pygmalion, placebo and other effects of expectation: some notes. University of Glasgow. Accessed 5 June, 2008 from <http://www.psy.gla.ac.uk/~steve/hawth.html>; last update 11 May 2008.
- [11] Oakley E, Stocker S, Staubli G, Young S. Using video recording to identify management errors in pediatric trauma resuscitation. *Pediatrics* 2006; 117(3): 658-64.
- [12] Warren J, Noone J, Smith B, Ruffin R, Frith P, van der Zwaag B-J, Beliakov G, Frankel H, McElroy H. Automated attention flags in chronic disease care planning. *Med J Aust* 2001; 175(6):308-12.
- [13] Bollen C, Warren J, Whenan G. Introduction of electronic prescribing in an aged care facility. *Australian Family Physician* 2005; 32(4):283-6.
- [14] Gadzhanova S, Iankov I, Warren J, Stanek J, Misan G, Baig Z, Ponte L. Developing high-specificity anti-hypertensive alerts by therapeutic state analysis of electronic prescribing records. *J Am Med Inform Assoc* 2007; 14(1):100-9.
- [15] Goodyear-Smith F, Kearse N, Warren J, Arrol, B. Trial of electronic medical textbooks DynaMed, MDConsult and UpToDate. *Australian Family Physician*, to appear.
- [16] Berry D, Hart A. Evaluating expert systems. *Expert Systems* 1990; 7(4):199-208.
- [17] Wharton C, Rieman J, Lewis C, Polson P. The cognitive walkthrough: a practitioner's guide. In *Usability Inspection Methods*. John Wiley, New York, 1994.
- [18] Dolin R, Alschuler L, Boyer S, Beebe C, Behlen FM, Biron PV, Shabo A. HL7 clinical document architecture, release 2. *J Am Med Inform Assoc* 2006; 13(1):30-9.
- [19] Leslie H. International developments in openEHR archetypes and templates. *HIM J* 2008;37(1):38-9.