Lessons Learned in Improving the Adoption of a Real-time
NLP Decision Support System

Yang Huang, PhD1, Daniel Zisook, MD1, Yunan Chen, PhD2,
Michael Selter1, Paul Minardi, MD1, John Mattison, MD1

1Kaiser Permanente Southern California
Pasadena, CA, USA
{yang.x.huang,daniel.s.zisook,michael.selter,
paul.m.minardi, john.e.mattison}@kp.org

2Department of Informatics
University of California, Irvine
Irvine, CA, USA
yunanc@ics.uci.edu

Abstract—While most research in the NLP domain focuses on information accuracy, the adoption of NLP applications in healthcare extends beyond technical innovations. This study investigates the adoption issues of an NLP application in three different field sites. Using both quantitative log analysis and qualitative user interviews, we identified four main factors that affect NLP adoption: organizational culture and support, system usability, information quality and system reliability. These factors must be considered to ensure successful adoption of NLP applications that provide real-time decision support in a clinical care setting.

Real-time; NLP; Decision Support; User Adoption

I. INTRODUCTION

Electronic Medical Record (EMR) systems have been used by clinicians to improve healthcare service for decades [1], yet significant barriers need to be overcome [2] to realize their full potential [3]. One of the barriers is that free text is hard to compute, yet its user-friendliness and superior expressivity make it the primary data format to record clinical information. Natural Language Processing (NLP) is a powerful technique used to extract computable information from unstructured clinical documents [4] such as radiology reports [5-7], mammography [8, 9], pathology reports [10, 11], and discharge summaries [12, 13]. The extracted data can then be used to enhance computerized clinical decision support in clinical domains [14-16]. Though much allure exists about the possibility of leveraging real-time NLP to provide decision support at the point-of-care, experience on successfully integrating such applications into an EMR and a physician workflow is sparse.

Enabling real-time clinical decision support from manually entered free text documents requires highly accurate abstraction of clinical information from ambiguous and fragmented text. Historically, medical informatics literature considers NLP applications as a technology-driven problem, focusing on intrinsic factors (“precision” and “recall”) as the prime determinant of adoption and success. Anecdotal data suggests that technological success may not always guarantee the use of these applications among physicians. Understanding the factors and processes that contribute to a successful NLP application adoption is important for achieving full benefit from such systems.

In the healthcare domain where clinical workflows, organizational hierarchies, and users’ cultural backgrounds are complex and intertwined, the human-computer interaction (HCI) poses a challenge to the adoption process. Many adoption studies focus on the use of large-scale IT systems such as CPOE or EMR systems but rarely focus on individual features such as adoption of an NLP coding decision support system.

Previous work suggests that over 40% of Healthcare IT implementations fail and lead to delayed system deployment, cost overruns, and even project abandonment [17-19]. Analyses of previous implementation failures have identified factors such as lack of clinicians’ involvement, poor physician leadership, and the concerns of drops in productivity and efficiency as contributors to adoption failure/resistance [20]. In order to avoid costly failures, it is important to consider these types of risks.

Few publications exist regarding factors affecting adoption of NLP applications in clinical systems. Mestre and Haug developed an Automated Problem List system composed of a background NLP problem extraction application and a foreground problem list management application [21]. It was integrated with an EMR to provide decision support for problem list management. The study focused on the technical design and accuracy of several NLP components and briefly mentioned application accuracy, speed, and coverage as potential issues of the system. Subsequent studies addressed the speed of the system, precision, recall and sensitivity [22, 23], but did not address other factors impacting system adoption.

Furthermore, the NLP proposed problems were added to the
problem lists within 2 days; therefore the NLP back-end system was not used as a real-time application.

To our best knowledge, this is the first published study assessing the adoption of a real-time NLP-based decision support system integrated in clinicians’ daily patient care workflow. The lessons learned from piloting a Natural Language Processing (NLP) application within an EMR system are presented to enhance future implementation of similar applications.

II. BACKGROUND

During the course of the last several years, Kaiser Permanente Southern California (KPSC) has gained significant insight into user factors contributing to a successful implementation of a real-time, point of care decision support application.

In 2006, KPSC began requiring all physicians to provide an appropriate Evaluation and Management CPT Code (E&M Code) per 1997 CMS E&M Coding Guidelines for all office visits. These codes are utilized to generate claims for physician services within the healthcare industry. As a significant portion of the membership dues at KPSC are from pre-paid membership and do not require E&M Codes, physician resistance to coding every encounter for billing purposes was much more pronounced than in similar healthcare environments that still operate on a predominantly fee-for-service model (FFS). In order to support this requirement, a project was launched to enable coding decision support to minimize the time required to generate an accurate and reproducible E&M code for each encounter. Calculating the E&M code for the most common office visit, a problem-based visit, requires the clinician to account for scoring the History of Present Illness (HPI), Physical Examination (PE), and Medical Decision Making (MDM) based on a defined set of criteria. It quickly became evident that to provide accurate decision support, users either would need to enter their notes as structured data directly into the EMR, or post-entry NLP would be required to transform narrative free text to into computable data.

While partnering with an NLP vendor to design an NLP system capable of accurately determining the components of an E&M code, KPSC devised several manual input methods for enabling the coding decision support. Figure 1 above depicts a form for inputting the narrative “history of present illness”. It contains check boxes representing the components needed to score the History portion of the E&M Code. The user is required to accurately determine which elements of the History are addressed in the free text.

Figure 2 below depicts a form for inputting the free text narrative of the physical exam. It contains a drop down menu for the user to input the total number of bullets addressed. The user must have the E&M coding knowledge to determine which bullets have been addressed in free text.

During the typical physician workflow, the physician begins documenting in the EMR after reviewing patient information including past labs, progress notes, allergies, pertinent medical history, nursing notes, and vitals. The documentation can be done using a combination of free text and structured data entry. The EMR vendor provides distinct structured form tools for inputting review of systems and physical exams. The user is not required to utilize these forms and can input the information as free text if preferred. While there is a preference for using form-based entry whenever appropriate and efficient, we sought to support the full expressivity of free text narrative without losing our capacity to automate the coding process in a reliable, reproducible, efficient, and compliant manner. In the outpatient setting the contents of the history of present illness are always entered as free text. The NLP engine is actively invoked by the physician when the free text portion of the note is complete using an “Auto Analyze” button shown in Figure 3.

Figure 4 below displays the results screen after NLP function has been invoked by the user. The user is required to review, edit as needed, and accept the results.
III. METHODS

In this study we employed both qualitative and quantitative methods to assess the adoption of this NLP decision support application among physicians in three different offices located in Southern California. System performance and user response to NLP results were logged. Continuous data collection was initiated at the onset of the study. Intermittent interviews with pilot physicians were conducted to coincide with major changes. These approaches allowed us to capture both the objective user interaction with the system and their subjective perceptions regarding the NLP applications in the clinical workflow. The data abstracted from the log allows comparison of how system efficiency, information accuracy, and both individual and organizational behaviors impact the actual use of the systems.

The system’s coding accuracy was evaluated using precision and recall at each coding evidence level (e.g. recording the duration of the addressed problem in History). Precision was defined as the percentage of correctly assigned coding evidences of all evidences assigned by the NLP engine. Recall was defined as the percentage of correctly assigned coding evidence of all the evidence in the gold standard and was constructed using three rounds of manual review by multiple auditors.

A. Pilot Sites

The initial pilot version of the NLP engine was completed on June 30, 2008. The analysis of a test corpus revealed a precision/recall of 88%/85%. The pilot users were identified and informed that a new button would be added to the Physical Exam form that would send the narrative note to the NLP engine and decipher the number of bullets addressed in their physical exam.

Three geographically distinct adult primary care pilot sites were identified. Medical Office Building (MOB) A included 15 physicians, MOB B included 16 physicians, and MOB C included 24 physicians. MOB A consisted of a group of young, technologically savvy physicians, while sites 2 and 3 had diverse physician ages and technical proficiency.

Over the next one and a half years, routine enhancements were implemented. On November 16, 2009, the application was rolled out to more than 300 adult primary care (APC) physicians in the San Diego area, and on December 14, 2009, it was rolled out to more than 1,500 APC providers in KPSC.

<table>
<thead>
<tr>
<th>Date</th>
<th>Event Description</th>
<th>Accuracy (Recall/Precision)</th>
<th>Speed (average time in seconds)</th>
<th>Average % of Encounters Analyzed by NLP (week prior/week after)</th>
<th>Average % of Users Utilizing NLP (week prior/week after)</th>
</tr>
</thead>
<tbody>
<tr>
<td>6/30/2008</td>
<td>NLP “auto-analyze” feature enabled at all pilot sites with text documentation forms.</td>
<td>PE.65%/88%</td>
<td>HX: NA, PE: 4 sec</td>
<td>A:0% /32%</td>
<td>A:0% /67%</td>
</tr>
<tr>
<td>10/11/2008</td>
<td>Accuracy Improvement</td>
<td>PE.66%/91%</td>
<td>HX: NA, PE: 4 sec</td>
<td>A:17% /15%</td>
<td>A:20% /33%</td>
</tr>
<tr>
<td>4/6/2009</td>
<td>NLP Application enabled for first time. Write notice was sent to pilot physicians 3 days prior to implementation.</td>
<td>PE.81%/96%</td>
<td>HX:4 sec, PE: 4 sec</td>
<td>C:19% /28%</td>
<td>C:21% /28%</td>
</tr>
<tr>
<td>4/13/2009</td>
<td>Onsite demo of new HIP functionality given at MOB C.</td>
<td>PE.86%/98%</td>
<td>HX:2.5 sec, PE: 2.5 sec</td>
<td>A:22% /39%</td>
<td>A:79% /64%</td>
</tr>
<tr>
<td>4/20/2009</td>
<td>Onsite demo of new HIP functionality given at MOB C.</td>
<td>PE.86%/98%</td>
<td>HX:2.5 sec, PE: 2.5 sec</td>
<td>A:19% /29%</td>
<td>A:64% /55%</td>
</tr>
<tr>
<td>5/16/2009</td>
<td>Onsite demo of new HIP functionality given at MOB C.</td>
<td>PE.86%/98%</td>
<td>HX:2.5 sec, PE: 2.5 sec</td>
<td>A:17% /15%</td>
<td>A:67% /60%</td>
</tr>
<tr>
<td>5/13/2009</td>
<td>Accuracy of HIP increased. BOS processing from HIP free text enabled.</td>
<td>PE.86%/98%</td>
<td>HX:2.5 sec, PE: 2.5 sec</td>
<td>A:17% /15%</td>
<td>A:67% /60%</td>
</tr>
<tr>
<td>8/5/2009</td>
<td>HIP and free text GUI components combined into a single &quot;Integrated&quot; form. Significant improvements implemented.</td>
<td>PE.86%/98%</td>
<td>HX:2.5 sec, PE: 2.5 sec</td>
<td>A:17% /15%</td>
<td>A:67% /60%</td>
</tr>
<tr>
<td>9/13/2009</td>
<td>HIP Accuracy Increase</td>
<td>PE.86%/98%</td>
<td>HX:2.5 sec, PE: 2.5 sec</td>
<td>A:17% /15%</td>
<td>A:67% /60%</td>
</tr>
<tr>
<td>10/14/2009</td>
<td>HIP Accuracy Increase</td>
<td>PE.86%/98%</td>
<td>HX:2.5 sec, PE: 2.5 sec</td>
<td>A:17% /15%</td>
<td>A:67% /60%</td>
</tr>
</tbody>
</table>
IV. RESULTS

During the pilot, numerous routine sequential enhancements were implemented. Table 1 represents a summary of key events. Important visual components were altered based upon physician feedback to improve the usability of the application. The ability to resubmit a note to NLP within the results screen, along with new functionality to show the specific text in the note where credit was given for a given bullet, were all very well-received. We analyzed the utilization data around important events to gain insight on how different factors impact user adoption. We define provider weekly adoption rate as the proportion of all providers using the NLP application at least once during the week divided by the proportion of all providers who had office encounters the same week. Encounter weekly adoption rate is defined as number of encounters (History or PE section) with NLP submissions divided by the total number of eligible office encounters with E&M codes recorded during the week.

A. Initial Deployment Usages

Figure 5 shows the NLP utilization rate for physical examination coding support by physician and encounter in three pilot sites, covering the two weeks prior and the six weeks after the application go-live on June 30th, 2008. Both MOB A and C experienced larger declines in provider utilization rate after the initial launch.

B. System Usage Changes in Response to Feature Modications

To increase the use of these NLP tools among physicians, several system-level modifications were made after the initial system deployment. Physicians’ acceptance rates were documented around the major modifications and show a steady improvement.

The key events that occurred during this time span include improvements in response time, functionality, and accuracy. In conjunction with on-site demos, the utilization rate in history section increased in all three sites. Figure 6 shows weekly total NLP adoption rates on Physical Examination and History by physician and encounter across all three sites.

C. Usage rate increase corresponding to Workflow improvement

Figure 7 below shows NLP adoption rates on Physical Examination and History by physician and encounter across all three sites. On August 30, 2009, a new integrated progress note box replaced two separate History and PE text boxes, streamlining the workflow. The average transaction time was further reduced to 1.6 seconds from 2.5 seconds. Subsequent to on-site demos at all three sites, a significant usage increase in both MOB A and MOB C was observed and this increase persisted in the following 10 weeks.
infrequent users who might have entered or dropped out of
user pool for reasons other than perceived coding accuracy.
These 3 periods were chosen because we did not have other
events such as on-site visits etc. between 8/30-10/31/2009
which might otherwise have significantly impacted the
utilization rate. Comparing two weeks after 9/13 with that
after 8/30, the number of physicians increasing their use of
NLP versus those decreasing usage was 18:9, with an
average of increase of 5.7%. When comparing periods of
10/14 and 9/13, the number of physicians increasing their
use of NLP versus those decreasing was 15:12, with an
average of increase of 3.8%. There is noticeable increase in
NLP use among NLP frequent users.

To assess the impact of speed improvement on user
adoption, we compared the utilization rate changes for two
releases on 10/11/2008 and 04/13/2009. The two releases
had identical PE functionality and accuracy, but the average
to-end response time has been reduced to 2.5 seconds
from 4 seconds. Among 11 PE NLP users during the two
weeks following 10/11/2008, 3 stopped using it 3 weeks
later. 5 of the remaining 8 users had decreased utilization
rates. The average utilization rate of all 11 users dropped 4%.
Among 9 PE NLP users during the two weeks following
4/13/2009, 1 stopped using it 3 weeks later and 4 had
decreased utilization rates. The average decrease was 1.3%,
which was more moderate than the 10/11/2008 release that
had a significant slower response time. In subsequent on-site
visits, several physicians disclosed that the response time
reduction was an important usability improvement that was
carried over to the History NLP application as well.

V. DISCUSSION

A. Organizational Culture and Support

Similar to what has been reported in other healthcare IT
adoption studies [20], organizational culture has greatly
impacted the adoptions of the NLP tools, as indicated by the
utilization rates in the three field sites. For instance, MOB A
had approximately 75% of eligible encounters analyzed by
NLP, while MOB B had approximately 33% of eligible
encounters being processed by NLP.

Much of this variation can be explained by the
organizational cultures in these three sites. Physicians at
MOB A are younger and thus more receptive to new
technologies. During site visits, it was noted that the
physician group at MOB A was more cohesive and
collaborative than at the other two sites. In addition, MOB A
is led by a young medical director who is keen on improving
patient care through technologies. Local physician
leadership is a critical success factor that may not overcome
significant technical barriers, but becomes absolutely critical
in optimizing effective adoption of any new technology.
This factor is increasingly relevant in a healthcare under
siege with major demands and initiatives from new
regulatory and legislative mandates and rapidly changing
market forces. The “change capacity” of most healthcare
organizations is already at a critical peak, so leadership is
critical for success in the following objectives: a) prioritization of work within a larger change management
framework, b) motivation to change, and c) optimization of
the process and the results of that change using true
outcomes-based metrics. The difference in organization
culture could also explain why MOB A had a larger % of
users (67%) versus the other 2 sites (44%, 47%) when the
pilot initially started. This trend was also demonstrated
when major enchantments in workflow were introduced
(e.g., History processing enabled, History/PE combined into
singe data entry form).

The impacts from in-person on-site demonstrations/visits
are also profound. Compared to written communications
such as email and flyers, this method got most physicians to
try the new feature more quickly. Information overload has
come a constant feature of the healthcare landscape.
Managing communication within that reality increasingly
requires explicit planning around communication strategies
that include message, target audience, modality, frequency,
and messenger.

After a successful initial launch, three other aspects are
critical to ensure a system has staying power among users:
usability, quality, and reliability.

B. Ease of Use and Usability of the NLP tool

Ease of use is crucial in interactive system design. It is
important to ensure integration into the existing physician
workflow, respecting the time-constraints that physicians
may have.

As shown in the results, NLP processing of free text PE
notes was initially uncommon. To address this issue, an
integrated text box which simplified users' submission was
created, allowing users to submit PE free text without
interrupting their workflow. This streamlining produced
notable increases in utilization rates for PE, particularly in
MOB C where there was a two-fold increase in PE Notes
processing after the integrated text box went live. This
increase may be partially attributable, however, to use of the
text box prior to the go-live date (August 30th), resulting in
later failure to detect PE text and an artificially low PE
utilization rate. This caveat underscores the importance of
workflow design and ensuring proper tool usage.

Concerns about system usability are also reflected in our
log analysis, as application speed and response time are
directly correlated with the use of the NLP tools. Several
physicians cite the reduction of average response time from 4
seconds to 2.5 seconds as impetus for increased usage of the
application. By reducing response time, expanding
functionality, and improving accuracy and on-site
demonstrations, the utilization rate in History section
increased in all three sites three months after initial History
rollout, a drastic comparison with the sharp drop-off after PE
roll out shown in Figure 5.

C. Information Quality

A quality system needs to provide users high-quality
information to be perceived as helpful versus distracting. As
our results show, physicians increased their use of the
application as its coding precision and recall were improved.
Text-highlighting was added to the application on 8/30/2009
allowing physicians not only see what bullets each note gets,
but also to view the specific words/phrases attributed to each bullet. This proved to be a very popular feature to improve user’s experience with the application.

VI. CONCLUSION

The value of NLP applications extend to general clinical practice, billing, and medical safety. The development and deployment of such tools often require huge financial, intellectual, and time investments, and they draw from a finite supply of cultural capacity and readiness to change. The benefits of these NLP tools cannot be realized unless they are successfully adopted by users. The present study analyzed the implementation of an NLP tool in three clinical environments in order to identify factors that may accounts for the variation in adoption rates across these different sites. Adoption was influenced by the quality of information it provided, but more importantly, by the users’ iterative and interactive design of the system, the organizational culture, and change management support. Factors such as information accuracy, usability, computer literacy, communication strategy, and a supportive physician lead are central to successful implementation. These findings support meticulous attention to each of these critical success factors for deployment of new systems in general and for guiding future NLP applications in particular.

ACKNOWLEDGMENT

We sincerely thank Sunil Karumuri for his tremendous help in developing the logging/reporting mechanism which allowed this study to happen. We also thank Brenda Goldstein for facilitating the on-site demonstrations and initial survey and Dr. Lee Ewin for helpful discussions.

REFERENCES


