Improving Completeness of Electronic Problem Lists through Clinical Decision Support: A Randomized, Controlled Trial

Citation

Published Version
doi:10.1136/amiajnl-2011-000521

Permanent link
http://nrs.harvard.edu/urn-3:HUL.InstRepos:10121116

Terms of Use
This article was downloaded from Harvard University’s DASH repository, and is made available under the terms and conditions applicable to Other Posted Material, as set forth at http://nrs.harvard.edu/urn-3:HUL.InstRepos:dash.current.terms-of-use#LAA

Share Your Story
The Harvard community has made this article openly available. Please share how this access benefits you. Submit a story.

Accessibility
Improving completeness of electronic problem lists through clinical decision support: a randomized, controlled trial

Adam Wright,1,2 Justine Pang,1,2 Joshua C Feblowitz,1,2 Francine L Maloney,2 Allison R Wilcox,2 Karen Sax McLoughlin,1 Harley Ramelson,1,2,3 Louise Schneider,1,3 David W Bates1,2,3,4

ABSTRACT
Background Accurate clinical problem lists are critical for patient care, clinical decision support, population reporting, quality improvement, and research. However, problem lists are often incomplete or out of date. Objective To determine whether a clinical alerting system, which uses inference rules to notify providers of undocumented problems, improves problem list documentation.

Study Design and Methods Inference rules for 17 conditions were constructed and an electronic health record-based intervention was evaluated to improve problem documentation. A cluster randomized trial was conducted of 11 participating clinics affiliated with a large academic medical center, totaling 28 primary care clinical areas, with 14 receiving the intervention and 14 as controls. The intervention was a clinical alert directed to the provider that suggested adding a problem to the electronic problem list based on inference rules. The primary outcome measure was acceptance of the alert. The number of study problems added in each arm was assessed. Data were collected during 6-month pre-intervention (11/2009–5/2010) and intervention (5/2010–11/2010) periods.

Results 17 043 alerts were presented, of which 41.1% were accepted. In the intervention arm, providers documented significantly more study problems (adjusted OR=3.4, p<0.001), with an absolute difference of 6277 additional problems. In the intervention group, 70.4% of all study problems were added via the problem list alerts. Significant increases in problem notation were observed for 13 of 17 conditions.

Conclusion Problem inference alerts significantly increase notation of important patient problems in primary care, which in turn has the potential to facilitate quality improvement.

Trial Registration ClinicalTrials.gov: NCT01105923.

INTRODUCTION AND BACKGROUND
An accurate and up-to-date patient problem list represents the cornerstone of the problem-oriented medical record, especially in internal medicine. It serves as a valuable tool for providers attempting to familiarize themselves with a patient’s clinical status and provides a means of succinctly communicating this information between providers. In addition, an accurate problem list has been associated with higher-quality care.1 For example, Hartung et al found that patients with ‘congestive heart failure’ (CHF) on their problem list were more likely to receive ACE inhibitors or angiotensin-II receptor blockers than CHF patients without ‘CHF’ listed on their problem list. Further, many clinical decision support (CDS) rules use problem list entries to make inferences about patients,2 so a complete, accurate list may facilitate more effective CDS. Conversely, an incomplete or inaccurate problem list could lead to delayed or inappropriate care. Finally, an accurate and comprehensive problem list would help to correctly identify patient populations and create patient registries conducing of quality improvement activities and research.

Despite these numerous benefits, problems lists are often inaccurate, incomplete, and out of date.3–5 In previous research, we showed that problem list completeness in one network ranged from 4.7% for renal insufficiency or failure to 50.7% for hypertension, 61.9% for diabetes, to a maximum of 78.5% for breast cancer,6 and other institutions have found similar results.3–5 In addition, we have found in previous qualitative studies that provider attitudes toward, and use of, the problem list vary widely.7 8

Beginning in 2011, in order to be considered ‘meaningful users’ of an electronic health record (EHR) and qualify to receive federal stimulus grants under the HITECH Act, which can total US $44 000 through Medicare and US$63 750 through Medicaid, providers must, among other things, ‘maintain an up-to-date problem list of current and active diagnoses,’ with 80% of patients having at least one problem recorded or an indication of ‘no known problems.’9–11 Given wide variation in problem list use by providers,7 8 new tools are needed to help providers meet this goal.

Researchers have used a variety of strategies in an attempt to detect patient problems and increase problem list use. In general, these methods fall into two broad categories: problem inference (or proxy) rules and natural language processing (NLP) techniques. Problem inference techniques use related clinical information such as laboratory tests, medications, and billing codes to infer problems (eg, a patient receiving metformin who has had multiple abnormal HbA1c tests is likely to have diabetes). In contrast, NLP strategies use algorithms designed to process and code free-text entries such as progress notes. Several groups have used data mining techniques and clinical associations to
predict patient problems. Others have reported success using NLP techniques to automate the problem list. Prior efforts have generally been evaluated in a laboratory setting, and focused on a single or small number of problems.

In this study, we performed a cluster randomized, controlled trial of a clinical alerting system that used inference rules to detect and notify providers of undocumented problems, giving them the opportunity to correct these gaps and improve problem list completeness. Our goal was to assess whether or not this system would improve problem notation for a broad array of patient conditions.

**METHODS**

**Design overview**

In a prior study, we presented a novel method for developing and validating problem-inference rules, as well as a knowledge base containing validated rules for 17 clinically important conditions (henceforth referred to as ‘study problems’). These rules were developed based on previous work using data-mining techniques to identify medicare-problem associations and laboratory-problem associations. The rules take into account account problem list entries (free-text and coded), billing diagnosis codes, laboratory results, medications, and vital signs to identify likely gaps in the problem list. Rule development and validation is described in detail in our previous work. To summarize, rule development occurred in six steps: (1) identification of problem associations with structured data; (2) selection of specific problems; (3) development of preliminary rules; (4) characterization of preliminary rules and alternatives; (5) selection of a final rule; and (6) validation of the final rule. Using these rules, the average sensitivity and positive predictive value (PPV) for the training set were 83.4% and 91.1%, respectively; for the validation set, average sensitivity and PPV were comparable at 83.9% and 91.7%, respectively. Importantly, the inference rules were more sensitive than the problem list itself and had a higher PPV than billing codes. The performance of the rules is fully described in our prior paper, and the contents of the rules are presented in the online appendix. As we developed the rules, we prioritized PPV and specificity in order to minimize the occurrence of false positive alerts, which might annoy users; however, for most conditions, we were able to achieve good performance on all four metrics: PPV, negative predictive value, specificity and sensitivity.

In four cases, we developed rules for groups of clinically similar entities: asthma/chronic obstructive pulmonary disease (COPD), congenital coagulopathy (hemophilia, congenital factor XI deficiency and von Willebrand disorder), osteoporosis/osteopenia, and renal failure/insufficiency. We created these groupings because, although we were able to determine, with a high degree of certainty, that the patient had one of the conditions (eg, asthma or COPD), we could not reliably discriminate between the conditions because of similar diagnostic criteria or treatment approaches. The 17 rules developed were:

- Attention deficit hyperactivity disorder
- Asthma/COPD
- Breast cancer
- Coronary artery disease (CAD)
- Congenital coagulopathy (hemophilia, congenital factor XI deficiency and von Willebrand disorder)
- CHF
- Diabetes mellitus
- Glaucma
- Hypertension
- Hyperthyroidism
suppressed for the duration of the study). When the provider adds a problem, he or she is also given the opportunity to add additional details or select a related term (eg, ‘gestational diabetes’ or ‘diabetes mellitus type 2’ instead of simply ‘diabetes mellitus’).

We conducted a randomized controlled trial of this intervention for a 6-month period, and also collected baseline data before the intervention in order to provide a second control. To reduce the risk of contamination, we used a cluster randomization method.

Clusters (n=28) were designated on the basis of pre-existing administrative divisions within the clinics. For example, one primary care clinic is divided into adult medicine, family medicine, and pediatric medicine, and another is divided into separate suites, A, B, and C. In both cases, these subunits were treated as separate clusters. Clusters were then grouped into three bands: hospital-based, community and federally qualified health center. Once grouped into the three bands, the clusters within each band were randomly allocated to the control or intervention arms, with 14 clinics randomized to the control arm and 14 to the intervention arm.

Providers were not aware to which arm their subclinic group was assigned until the intervention was implemented. Patients were not made aware of the intervention. No pre-intervention orientation or training took place in the intervention arm. Blinding was not possible given the nature of this intervention. No pre-intervention data were not made aware of the intervention. No pre-intervention data were not made aware of the intervention.

For the primary outcome, we calculated the acceptance rate of the alert for each of the 17 conditions, as well as an overall acceptance rate. For the secondary outcome of problem addition, which consisted of comparisons of count data, we modeled our data as Poisson-distributed counts. The unadjusted relative rate was calculated as described above, and tested for equality with one using a normal approximation.

In addition to this unadjusted relative rate, we used Poisson regression with an interrupted time series approach to control for potential exogenous temporal effects. Specifically, we used five coefficients and a scale parameter to model six features: starting rate, four slopes (pre and post period for the control and intervention arms), and a parameter for effect of the intervention. The effect parameter was an OR for the immediate effect of the intervention. In the case where differences between the control and intervention groups were non-significant, we removed the related terms from our model. This resulted in a new parameter for the intervention, which instead measured the overall effect of the intervention. This parameter has a similar interpretation to our unadjusted relative rate, and was compared for equality with one using a $\chi^2$ test.

Finally, in order to counteract a possible problem of multiple comparisons, we used a Bonferroni correction. This correction maintains the error rate by testing each hypothesis against a lower $\alpha$ value, where the new cut-off for statistical significance is $\alpha/n$, where $n$ is the number of independent tests. In our case, the new cut-off was calculated to be 0.0029 (0.05/17 rules).

Demographic data were analyzed using a $\chi^2$ test for categorical data and Student t test for continuous variables. Study data were analyzed using SAS V9.2.
Role of the funding source
This work was supported by a grant from the Partners Community HealthCare Incorporated System Improvement Grant Program. Partners Community HealthCare Incorporated was not involved in the design, execution or analysis of the study, or in the preparation of the manuscript.

RESULTS
Participant flow
All 28 participating clinics completed the study, and there was no loss to follow-up. Overall, 41,039 patients were seen in the control clinics during the entire study period, and 38,025 patients were seen in the intervention clinics. A small number of patients (n = 3,894, 5.2%) were seen in both intervention and control clinics, and thus appear in both arms of the study. Figure 2 shows the flow of subclinics through the study.

Demographic and baseline data
Intervention and control groups appeared clinically similar across a range of demographic and clinical variables (Table 1). During the 6-month pre-intervention period, greater problem list use was observed in the control group. A total of 3,230 study problems (17.8 problems/day) were added in the intervention group, and 3,597 study problems (19.8 problems/day) were added in the control group (p < 0.001).

Primary outcome: acceptance rate
Problem inference rules fired a total of 17,043 times during the intervention period for a total of 11,508 patients in the intervention arm. The overall acceptance rate for problem inference alerts was 41.1%. The highest acceptance rate of the 17 conditions was 55.7% for glaucoma alerts (Table 2). Alerts for myasthenia gravis and sickle cell disease were infrequently presented and infrequently accepted.

Pre-specified secondary outcome: problem notation in the problem list
During the intervention period, 10,016 study problems were added in the intervention group compared with 3,759 added in the control group—an absolute difference of 6,277 problems (compared with 367 fewer problems added in the intervention group during the pre-intervention period, p < 0.0001). The unadjusted relative rate of study problem addition was 2.98 times more problem notation in the intervention group (p < 0.0001), and the adjusted OR was 3.43 (p < 0.0001).

The cumulative number of study problems added over the course of the entire study is shown in Figure 3. As reflected in the figure, the rate of study problem notation during the pre-intervention period was slightly lower in the intervention group than in the control group. The inflection point in the intervention group line was coincident with the initiation of the study intervention in that group and, by the completion of the study, the intervention group had added significantly more problems than the control group.

Table 3 shows the rate of problem list addition for each of the 17 study problems. Using the unadjusted differences measure, statistically significant increases in problem notation were seen for 15 of 17 study problems using an uncorrected threshold of p < 0.05. When the Bonferroni correction was applied to the threshold, two of the 15 problems were no longer statistically significant (congenital coagulopathy and hyperthyroidism). Relative rates of problem notation (for statistically significant conditions) ranged from 1.54 times more notation for hyperthyroidism (p = 0.031) to 6.89 times more notation for renal failure and insufficiency (p < 0.0001).

In addition to the unadjusted difference, we used Poisson regression and interrupted time series analysis to control for temporal trends. We began with a model with four slopes (results not shown). Outcomes from this model were similar to

Figure 2 Participant flow for study clusters (subclinic randomization). Because randomization was carried out at the subclinical level, a small number of patients (n = 3,849) appear in both arms of the study.

Table 1 Demographics of patients seen in control and intervention clinics

<table>
<thead>
<tr>
<th>Demographic characteristic</th>
<th>Control</th>
<th>Intervention</th>
<th>p Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Age (years), mean (SD)</td>
<td>49.6 (20.0)</td>
<td>47.7 (19.6)</td>
<td>&lt;0.0001</td>
</tr>
<tr>
<td>Sex (female)</td>
<td>61.4%</td>
<td>68.0%</td>
<td>&lt;0.0001</td>
</tr>
<tr>
<td>Race or ethnicity</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>American Indian</td>
<td>0.12%</td>
<td>0.17%</td>
<td>&lt;0.0001</td>
</tr>
<tr>
<td>Asian/Pacific Islander/Native Hawaiian</td>
<td>2.4%</td>
<td>3.0%</td>
<td></td>
</tr>
<tr>
<td>Black</td>
<td>14.2%</td>
<td>11.4%</td>
<td></td>
</tr>
<tr>
<td>Hispanic</td>
<td>17.1%</td>
<td>17.6%</td>
<td></td>
</tr>
<tr>
<td>White</td>
<td>57.6%</td>
<td>58.6%</td>
<td></td>
</tr>
<tr>
<td>Other</td>
<td>0.61%</td>
<td>0.55%</td>
<td></td>
</tr>
<tr>
<td>Declined/unknown</td>
<td>7.9%</td>
<td>8.7%</td>
<td></td>
</tr>
<tr>
<td>Language</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>English</td>
<td>85.7%</td>
<td>85.3%</td>
<td>&lt;0.0001</td>
</tr>
<tr>
<td>Spanish</td>
<td>8.8%</td>
<td>10.0%</td>
<td></td>
</tr>
<tr>
<td>Other</td>
<td>5.5%</td>
<td>4.6%</td>
<td></td>
</tr>
<tr>
<td>Primary insurance</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Commercial</td>
<td>59.6%</td>
<td>64.1%</td>
<td>&lt;0.0001</td>
</tr>
<tr>
<td>Medicare</td>
<td>22.3%</td>
<td>17.7%</td>
<td></td>
</tr>
<tr>
<td>Medicaid</td>
<td>14.0%</td>
<td>14.6%</td>
<td></td>
</tr>
<tr>
<td>Other/self pay</td>
<td>4.0%</td>
<td>3.6%</td>
<td></td>
</tr>
<tr>
<td>Income (US$), mean (SD)</td>
<td>56,350 (20,688)</td>
<td>59,663 (23,737)</td>
<td>&lt;0.0001</td>
</tr>
</tbody>
</table>
unadjusted results; however, the increases for congenital coagulopathy and hyperthyroidism were no longer statistically significant (at either $p < 0.05$ or the Bonferroni-corrected threshold of $p < 0.0029$). After removal of non-significant model components (yielding a simplified two-slope model shown on the right-hand side of table 3), the difference for congenital coagulopathy was once again statistically significant; however, when the Bonferroni correction was used, this study problem was not statistically significant. ORs from the final model were mostly similar to the unadjusted relative rates, and the overall OR for intervention effect on problem list notation was 3.43 ($p < 0.0001$).

To assess the accuracy of problems added as the result of the intervention, we also conducted an audit of a random selection of accepted alerts ($n = 1178$). In order to form a representative sample, we used a weighting strategy. Each of the 17 may have been suggested on the basis of one or more condition sets (eg, a diabetes suggestion could be triggered by the HbA1c value, medications, billing codes, or a combination of these features). For each condition set, we reviewed the accuracy of up to 30 alerts (less if there were fewer than 30 total accepted alerts for a given condition set). Study staff (FM) conducted a manual chart review, including a review of the notes. The gold standard was free-text documentation of the problem in any stored physician notes. We computed a weighted accuracy score by taking the accuracy for each condition set and weighting it according to how often that condition set triggered an alert. The weighted accuracy of all accepted alerts was found to be 89.8%. The 10.2% of alert acceptances not associated with documentation had a variety of causes, including patients near to diagnosis of a disease (eg, patients with pre-diabetes or metabolic syndrome on the cusp of diagnosis with diabetes), patients who appeared to actually meet diagnostic criteria for the disease, but for whom the diagnosis was not discussed in the record (eg, patients who met diagnostic criteria for chronic kidney disease based on glomerular filtration rate or hypertension based on serial blood pressure measures but without documentation of the condition in their notes), as well as some potentially erroneous additions.

**DISCUSSION**

We found that electronic problem list alerts were often accepted by users, and resulted in a substantial increase in study problem notation. The rate of notation of study problems increased dramatically during the intervention period as a result of this simple alert-based intervention. Overall, study problems were approximately three times more likely to be documented when alerts were shown. This increase is clinically important, since many of these problems are used for quality improvement and CDS.

Importantly, 14 out of 17 study problems were more often recorded in the intervention group than the control group. Only three conditions, myasthenia gravis, sickle cell disease, and hyperthyroidism, had similar rates between the two groups; however, even though the difference for hyperthyroidism was not statistically significant with Bonferroni correction, one could infer that there may be a trend for possible statistical significance with a larger sample size. Since our previous research validated the algorithm for the study problems, it is probable...
that the overall low prevalence of myasthenia gravis and sickle cell disease is responsible for the lack of any difference in notation between study arms.

Our results suggest that problem inference rules such as these are a valuable tool for improving problem list completeness and thus may be beneficial for improving patient care. A more complete problem list makes it easier for providers to obtain an accurate picture of a patient’s issues, which is especially important when an unfamiliar patient is being seen, such as in the case of urgent care or emergency visits, or in inpatient wards. Additionally, since problems are used for CDS, identification of patients for research studies, and quality measurement, these types of rules show great potential for improving quality and reducing costs.

One important question is how the observed increase in the notation of problems would ultimately benefit patients. Assuming that a given alert was correct, there were two potential scenarios for each alert reminder: (1) the alert called attention to an undocumented problem that the provider was not aware of and (2) the alert recommended a problem that the provider was aware of but had not documented in the problem list. Although the first scenario may provide a particular immediate clinical impact (making the provider aware of an unknown diagnosis), it is also likely to be less common. However, both cases provide significant positive clinical benefit, including enabling CDS (such as relevant preventive care reminders), facilitating quality measurement and research, and promoting awareness of a patient’s active problems among the entire care team (including providers that may not know the patient well).

An additional implication of this study may be to help providers achieve ‘meaningful use’ of EHRs, as one of the stage 1 and 2 meaningful use goals is to demonstrate problem list use for 80% of patients over the next few years.\(^9\)\(^{10}\) By meeting the meaningful use criteria, clinicians would receive incentive funds that could offset the expenses of implementing and maintaining the LMR. A tool such as that described here may be highly valuable for encouraging problem list use and increasing accuracy in the near term, especially for the large numbers of providers who are just starting with electronic records and are struggling to populate their problem lists.

Given these promising results and diverse potential applications, we hope to dramatically expand the problem inference knowledge base in the future. Ultimately, rules such as these may be used in tandem with provider documentation to increase the accuracy of the problem list, with the potential to improve patient care. However, additional provider engagement will also be required, and some problem list maintenance tasks (such as the removal of resolved problems or consolidation of duplicates) are beyond the scope of our described intervention.

### Limitations

Our investigation has several potential limitations. First, problem inference rules were developed, validated, and tested at a single site. Further research will need to be carried out to assess the generalizability of these results. Additionally, we had the benefit of a self-developed EHR, giving us the ability to extract the necessary data to develop and validate our knowledge base, as well as the ability to design a novel intervention. In contrast, most institutions use commercial EHR systems, which may not have this degree of flexibility. Although we encourage other institutions to develop and validate their own rules when feasible, we have also made our full knowledge base freely available for use by other organizations, including vendors.\(^5\) Another possible limitation of this approach is that imperfect accuracy of the problem inference rules could lead to erroneous alerts, which, if accepted, would result in inaccurate problems being added to the problem list. An audit of a random selection of accepted alerts revealed a global weighted accuracy of 89.8%. Although the accuracy of accepted alerts was very high overall, this finding nevertheless reveals the presence of a number of problems erroneously added to the problem list as a result of the alerts. Many of these instances appeared to be the result of borderline conditions (eg, metabolic syndrome, white-coat hypertension). However, the potential problem of providers accepting erroneous clinical alerts merits further study.
Additionally, the overall acceptance rate for the alerts was 41.1%, with the rest of the alerts being either over-ridden or ignored by physicians. Since the accuracy of the accepted alerts was shown to be high in the above-described audit, future research should look into the reasons why alerts are either over-ridden or ignored. Finally, the intervention was limited to primary care providers; extending this tool to specialties may require more focus on the types of alerts presented.

CONCLUSION
Problem inference alerts appear to be a powerful tool for improving notation of patient problems, and may thus in turn help improve quality of care. The use of problem inference alerts dramatically increased the notation of patient problems in the intervention group. Healthcare providers seeking to increase problem list completeness for meaningful use or other reasons should consider implementing such alerts. Future studies should focus on whether implementing such alerts has a direct effect on patient outcomes.

Acknowledgments We thank the following: Don Bugbee, Betti Rozenman, Regina Breyt, Janak Joshi and Bill Hanley for designing, developing and implementing the MAPLE software; Teal Aroy, Sue Smith, Janet Cygielnik, Dave Dubois, Kira Tsivkin Breyt, Janak Joshi and Bill Hanley for designing, developing and implementing the Blumenthal D, Tavenner M. The “meaningful use” regulation for electronic health records. N Engl J Med 2010;363:501–4.


