Presence of key findings in the medical record prior to a documented high-risk diagnosis

Mitchell J Feldman,1,2 Edward P Hoffer,1,2 G Octo Barnett,1,2 Richard J Kim,1 Kathleen T Famiglietti,1 Henry Chueh1,2

ABSTRACT

Background Failure or delay in diagnosis is a common preventable source of error. The authors sought to determine the frequency with which high-information clinical findings (HIFs) suggestive of a high-risk diagnosis (HRD) appear in the medical record before HRD documentation.

Methods A knowledge base from a diagnostic decision support system was used to identify HIFs for selected HRDs: lumbar disc disease, myocardial infarction, appendicitis, and colon, breast, lung, ovarian and bladder carcinomas. Two physicians reviewed at least 20 patient records retrieved from a research patient data registry for each of these eight HRDs and for age- and gender-compatible controls. Records were searched for HIFs in visit notes that were created before the HRD was established in the electronic record and in general medical visit notes for controls.

Results 25% of records reviewed (61/243) contained HIFs in notes before the HRD was established. The mean duration between HIFs first occurring in the record and time of diagnosis ranged from 19 days for breast cancer to 2 years for bladder cancer. In three of the eight HRDs, HIFs were much less likely in control patients without the HRD.

Conclusions In many records of patients with an HRD, HIFs were present before the HRD was established. Reasons for delay include non-compliance with recommended follow-up, unusual presentation of a disease, and system errors (eg, lack of laboratory follow-up). The presence of HIFs in clinical records suggests a potential role for the integration of diagnostic decision support into the clinical workflow to provide reminder alerts to improve the diagnostic focus.

INTRODUCTION

The Institute of Medicine’s pivotal reports of 1999 and 2001 documented the scope of potentially preventable errors in healthcare and suggested strategies for reducing these errors. Recent studies have explored preventable errors in the ambulatory setting through analysis of workflow and malpractice claims data. These studies have singled out diagnostic mistakes as a common, preventable source of error. In a study of ambulatory adverse events, investigators found that diagnostic errors led to 56% of preventable adverse events. Other studies of ambulatory malpractice claims showed that nearly half (48%) of the errors related to technical competence were due to failure to consider the correct diagnosis at the time of decision making, and over a third of diagnostic errors were due to incorrect interpretation of test results. A recent multihospital study documented diagnostic adverse events in 0.4% of all hospital admissions, and 83% were felt to be preventable. Failure to diagnose has been most studied in the specialties of family practice and internal medicine, but is a problem in the surgical specialties as well. Trainees (interns, residents, and fellows) are felt to be at particular risk of making diagnostic errors, and over one-fifth of malpractice claims involve trainee error. A computer-based intervention to reduce diagnostic error that is integrated into the workflow will be much more likely to be used than one that must be actively sought out. It will also probably require less time than a more traditional paper-based quality improvement intervention. As increasing numbers of physicians use electronic medical records, an integrated intervention may have a large impact on improving patient care.

In a recent British study, researchers found that, in women with ovarian cancer (one of the high-risk diagnoses (HRDs) from our study), symptoms such as abdominal distention, abdominal pain, and urinary frequency were noted to be present in the medical record weeks to months before the diagnosis. We believe that an alerting system that recognizes clinical clues and raises HRDs for consideration at an earlier stage could provide significant patient benefit.

Computer-based expert diagnostic systems exist, but their integration into clinical workflows has been limited. Before embarking on an intervention study to determine the effectiveness of a diagnostic decision support reminder system, however, we first need to document whether clues to an HRD exist in the clinical notes at a stage earlier than the documentation of that diagnosis. The principal objectives of this study were to (1) demonstrate that important, high-information clinical findings (HIFs) could be identified for selected HRDs, (2) manually review medical records to search for the presence of these HIFs before the documentation of the HRD, and (3) search for the same HIFs in a group of control patients without the index disease, to compare the incidence of finding HIFs in notes of patients with and without a known HRD. Ultimately, by alerting physicians to important diseases that might otherwise go unrecognized, and delivering this information while the patient is still under evaluation, patient care may be improved through timely consideration of other potential diagnoses.

Medical diagnostic decision support systems (MDDSSs) contain rich knowledge bases (KBs) of symptoms, signs, and laboratory findings, along
with probabilistic data linking these findings to the diseases in which they occur. Textbooks and journal articles often provide information about the frequency with which findings occur in diseases (eg, sensitivity), but rarely about the likelihood of the occurrence of a disease given the presence of a finding (eg, positive predictive value). MDDSS KBs are built using a variety of sources including review articles, meta-analyses, textbooks, and other knowledge resources. The sensitivity information is derived from the medical literature. The positive predictive value data are also derived from the medical literature where available, and from the judgment of physician editors when unavailable in the literature. Because MDDSSs contain quantitative information about both sensitivity and positive predictive value, we feel that they are a ready, generalizable resource for determining HIFs. One such MDDSS is DXplain, a mature web-based system developed and supported by the Laboratory of Computer Science (LCS) at the Massachusetts General Hospital (MGH). In a study by Berner et al., the performance of DXplain and other diagnostic decision support systems were compared on a set of 185 challenging cases. DXplain scored well on inclusion of the case diagnosis in its differential, comprehensiveness of both the KB and differential, and relevance of the differential. The authors found that DXplain suggested an additional two to three diagnoses per case that experts found relevant but that were not originally considered. We used the DXplain KB as a paradigm MDDSS for finding HIFs.

Others have reviewed medical records for a single disease looking for indicators of earlier detectability. Our study adds to this literature by reviewing clinical notes from patients with each of eight known HRDs to identify whether findings suggestive of these HRDs were present in the notes before establishment of the HRD, and compares these data with those from age- and gender-compatible controls.

The thrust of our study can be summarized by the following three questions and the flow diagram (figure 1):

1. Are the HIFs in the record?
2. Can we identify the HIFs, and are they found more commonly in patients with an HRD than in control patients without the HRD?
3. What was the time interval between the documentation of the HIFs and that of the HRD?

METHODS

Phase I: identifying important HIFs for selected HRDs

We determined HRDs to be those in which delayed diagnosis may result in significant morbidity or mortality and which are seen in the MGH ambulatory practices at least four times a year. Eight HRDs were selected: breast cancer, acute myocardial infarction, placement of intervertebral disc, lung cancer, appendicitis, colon and rectal cancer, bladder cancer, and ovarian cancer. The first six of these diseases are the most common conditions leading to medical malpractice claims in the USA. We used the numerical attributes that link findings to diseases in DXplain (crude probabilities of the finding’s frequency of occurrence in, and the positive predictive value for, the disease) to identify those terms that have high ‘information content’. For each HRD, based on the MDDSS KB, we identified findings that strongly suggest the presence of the disease (high ‘evoking strength’, which is a crude approximation of positive predictive value) and/or findings that occur very often (>80% of the time) in the disease (high ‘term frequency’, which is a crude approximation of sensitivity). We also identified findings that occur fairly often (~50–80% of the time) in the disease and that had high overall clinical significance (high ‘term importance’ value in DXplain).

Phase II: proof-of-concept using research patient data registry (RPDR)

LCS staff physicians (MF, EH) reviewed actual clinical problem lists, laboratory results, and notes from electronic health records for specific cases to identify HIFs extracted from the MDDSS HRD profiles in phase I. Rather than attempting to find all clinical terms, we concentrated our efforts on these HIFs for the eight HRDs. We retrieved patient records from the RPDR for each of the eight pre-specified HRDs. The query to the RPDR retrieved outpatient records from MGH where the HRD first appeared in the record after a cut-off date, chosen so that at least 20 records would be retrieved (January 1, 2005 for bladder cancer and acute myocardial infarction, January 1, 2008 for lung cancer, and January 1, 2006 for the other five HRDs). A total of 1379 patient records were retrieved using the database queries. Of these, 623 patients had the HRD appear on the problem list. From these 623 records, 243 were reviewed sequentially. We reviewed consecutive records from a convenience sample of at least 20 patients for each of the eight HRDs. For each of these patients, visit notes from before establishment of the HRD were reviewed. The reviewers inferred that a diagnosis had not yet been established if there was no mention of the diagnosis in the note, the diagnosis was not on the problem list, and the diagnosis was documented at a future visit. For each patient record reviewed, the reviewer recorded HIFs, if present, and the number of days from when HIFs first appeared in the patient’s record until the diagnosis was established.

For each set of patients with a given HRD, we used the same group of 623 records to find age- and gender-compatible control patients, sequentially, from within the other HRD records (control notes used were from patients with different HRDs from the index patient). Thus, if the index patient were a woman in her seventies with colon cancer, the control note might be from the medical record of a woman in her seventies with lumbar disc disease. Ideally, we found a ‘complete physical’ or ‘comprehensive visit’ note. However a ‘follow-up’ note or ‘subspecialty note’ could be used if it were a detailed note that...
included a physical examination. Also, if the control patient note indicated that this patient coincidentally had an active index disease, then another control patient was assigned. We reviewed each control note in the same manner as the index note, looking for the same HIFs of the index disease. We performed Fisher’s exact test to compare the frequency of occurrence of HIFs in records of patients with an HRD with the frequency of occurrence of HIFs in the control records.

**Human studies**

Institutional review board approval through MGH was obtained before the start of the study.

**RESULTS**

An average of 59 (range 13–90) HIFs per HRD were identified from the MDDSS disease profiles. An average of 30 (range 11–61) of the HIFs were selected on the basis of high evoking strength, and an average of 10 (range 1–29) findings did not have high evoking strength, but had a very high term frequency, or had a high term frequency and high overall clinical importance.

A total of 243 records were reviewed, at least 20 for each HRD. Overall, 25% of records reviewed (61/243) had HIFs present in the notes before the HRD was made. The mean number of days between the first appearance of HIFs in the patient’s record and establishment of the diagnosis varied greatly by HRD, ranging from 19 for breast cancer to 747 for bladder cancer (table 1).

Of records with HIFs, most had three or four HIFs each (range 1–8), except for breast and bladder cancer records; these had one each.

In three HRDs (lumbar disc disease and ovarian and bladder cancer), the number of control patient notes containing HIFs was significantly different from the number of index patient notes containing HIFs (Fisher’s exact test p < 0.05) (table 1). Figure 2 is a dot plot where each dot represents the number of HIFs found in a single patient record.

During the manual record review, we encountered many unconventional abbreviations (eg, AJ, KJ), nuanced textual descriptions (eg, ‘feels ‘tightening along the sides of her abdomen’), misspellings, and other semantic issues (table 2).

**DISCUSSION**

A quarter of records reviewed (61/243) contained HIFs in the record before establishment of the diagnosis. This range from 8% for the diagnosis of myocardial infarction to 55% for the diagnosis of lung cancer. There are several potential reasons why so many records did not contain HIFs (note the preponderance of dots in figure 2 at the 0 HIF level):

1. Many patients were referred to MGH with the diagnosis already known.
2. Many patients had been diagnosed in the remote past, but the diagnosis was not placed on the problem list until many months or even years later.
3. Many diagnoses were picked up incidentally during screening tests in asymptomatic patients. In these cases, one would not expect HIFs to be present before the diagnosis was made.
4. Some patients presented with signs or symptoms of an HRD but at prior visits were asymptomatic with no HIFs noted. This timing of occurrence probably explains why patients with acute myocardial infarction had the lowest incidence of HIFs noted in the record before the diagnosis was established.

Although most of the HRDs were made promptly once HIFs were noted in the records, even at a tertiary care academic medical center, there were still some cases where the HRD was not evident despite HIFs having been noted in the records. Thus, although it is reassuring that most HRDs are identified swiftly, we feel there is room for improvement, perhaps via diagnostic decision support, in a minority of cases. For example, for 55% of patients with colon cancer, the diagnosis was established within a few days (1–5) of HIF being first noted in the record. In 36%, the diagnosis was established within a few weeks, and in 9% over 1 year elapsed. For breast cancer, all but one patient was diagnosed within 10 days of HIFs appearing in the record; in the remaining patient, this occurred 2 months later. It is highly likely that diagnosing breast cancer 10 days earlier would not have produced a measurable long-term benefit to the patient.

Therefore, when considering which high-risk conditions to target for an intervention such as this, it will be important to prioritize resources to those conditions in which historically there has been the greatest delay, with the greatest chance for patient benefit. With ovarian cancer, 67% of patients had a diagnosis within 2 weeks of HIFs being noted; the longest delay was 3.5 months. In patients with bladder cancer, there was a much wider range in the time elapsed from HIFs to diagnosis: from 2 days to almost 10 years. In patients ultimately diagnosed as having lumbar disc disease, for 67% of them, the diagnosis was made within 1 month of HIFs appearing in the record. With the other 33%, the duration between HIFs appearing in the note and diagnosis ranged from 1 to 6 years.

It is likely that the prolonged duration from HIF documentation to diagnosis in some cases of lumbar disc disease was due to (A) overlap of clinical findings with less concerning diagnoses such as muscular low back pain, and (B) the accepted clinical practice of conservative therapy even in patients with

<table>
<thead>
<tr>
<th>Table 1 Results of record review for eight high-risk diagnoses</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>HRD</strong></td>
</tr>
<tr>
<td>---------------------------------------------------------------</td>
</tr>
<tr>
<td>Colon cancer</td>
</tr>
<tr>
<td>Breast cancer</td>
</tr>
<tr>
<td>Acute myocardial infarction</td>
</tr>
<tr>
<td>Lumbar disc disease</td>
</tr>
<tr>
<td>Lung cancer</td>
</tr>
<tr>
<td>Ovarian cancer</td>
</tr>
<tr>
<td>Bladder cancer</td>
</tr>
<tr>
<td>Appendicitis</td>
</tr>
</tbody>
</table>

Values in last column are mean ± SD (range). Dx, diagnosis; ES, evoking strength; HIF, high-information clinical finding; HRD, high-risk diagnosis; RPDR, research patient data registry; TF, term frequency.
dictionaries to include customized vocabularies such as those in standardized vocabularies such as the UMLS. Medical concepts from narrative text and mapping to related systems failure (eg, lack of laboratory follow-up).

The most common apparent reason for delay was patients not complying with recommended follow-up (eg, patients not complying with referral for colonoscopy with a positive guaiac test or with referral to urology in follow-up for hematuria). Other reasons include unusual presentations of a disease and systems failure (eg, lack of laboratory follow-up).

Medical students have long been taught that the contributions of history, physical examination, and laboratory results to a patient’s diagnosis are respectively 80%, 10%, and 10%. Studies including those by Hampton et al\textsuperscript{20} and Peterson et al\textsuperscript{21} support this age-old adage. Natural language processing (NLP) is necessary to extract the crucial history elements from narrative text to allow any automated form of diagnostic decision support.

Many decision support tools, including MDDSSs, require specific controlled vocabularies for input. Fortunately, many NLP systems in the clinical medicine domain, including MedLEE\textsuperscript{22} and cTAKES\textsuperscript{23} are evolving to permit extraction of medical concepts from narrative text and mapping to related concepts in standardized vocabularies such as the UMLS Metathesaurus. These systems also allow extension of their dictionaries to include customized vocabularies such as those used by decision support systems.

Automated extraction of clinical findings from a patient’s record is one important goal of diagnostic decision support. Our manual review illustrated some of the issues with which NLP experts grapple regularly, including unusual abbreviations, misspellings, and concepts expressed in uncommon ways (see table 2). The construction of large synonym dictionaries, possibly facilitated by multi-institutional collaborations and informed by the parsing of large clinical corpora, will probably be necessary to optimize the chances of mapping these highly variable lexical variants. Once automated extraction of HIFs is achieved, passing them from the clinical note to a diagnostic decision support reminder system may allow such a system to alert the clinician. Some examples from our review that we believe apply here are: an unusual presentation of appendicitis with a negative CT scan; a subtle presentation of ovarian cancer with initial symptoms more suggestive of a gastrointestinal problem; a patient ultimately diagnosed as having bladder cancer who had hematuria noted 10 years previously and a urine cytology report suspicious of malignant cells 9 years previously. The latter example must have been an error in laboratory follow-up, but perhaps a reminder system would have brought the diagnosis to consideration earlier. In the former examples, it is also possible that a reminder about an HRD by a diagnostic decision support system may have triggered the clinician to consider this in the differential, allowing earlier diagnosis and treatment.\textsuperscript{24}

We used control notes to address the concern that the identification of these key HIFs in the records of patients with HRDs could have been by chance. In three of the HRDs, where the Fisher’s exact test comparing the proportion of index records with HIFs with the proportion of control records with HIFs was significant, the results supported that HIFs were unlikely to be present in the records of patients with an HRD by chance. In four other HRDs, the control notes contained HIFs often enough that the Fisher’s exact test result was not significant, suggesting that, for patients with these HRDs, the HIFs would not have been good predictors for the HRD. We believe that three reasons contribute to this phenomenon. First, while most of the HIFs selected by our algorithm had a high predictive value for the HRD, some of the HIFs chosen that were based solely on the frequency of occurrence of the HRD, also occur commonly in the general population. In hindsight, these were not ideal candidates for HIFs and should not have been included. Thus cigarette smoking, hyperlipidemia, and coronary artery disease occur often enough in the general population (and in our control notes) that their use as HIFs for the HRDs, lung carcinoma and acute myocardial infarction, was problematic. For example, ‘cigarette smoking’ was found in 40% of the control patient notes as the sole HIF, and in none of the patients with the actual HRD as the sole HIF (when present in patients with lung carcinoma—the actual HRD—there were always other HIFs noted). Second, in the index patients, we looked at the visit note before the diagnosis was established and this was often not a comprehensive visit note. However, for the control patients, we looked mostly at the complete annual physical examination note, which would be more likely to contain a full review of systems, history, and laboratory data that might be expected to (and did) document hyperlipidemia, coronary artery disease, etc. This certainly played a role with the HRD, acute myocardial infarction, where a significant Fisher’s exact test result was indicative of the number of control notes containing HIFs being much greater than the number of index notes containing HIFs. Third, for some HRDs, there were too few index patient records in which HIFs could be found to allow a significant Fisher’s exact test result. Thus for the HRD, breast cancer, although there were four index patient records with HIFs and zero control patient records with HIFs, the Fisher’s exact test only approached, but did not reach, significance (p=0.12). We imagine

### Table 2: Medical terminology extracted by human reviewers

<table>
<thead>
<tr>
<th>Text in note</th>
<th>HIF term</th>
</tr>
</thead>
<tbody>
<tr>
<td>‘RLQ pain’</td>
<td>Right lower quadrant abdominal pain</td>
</tr>
<tr>
<td>‘AJ &amp; KJ less on left side’</td>
<td>Ankle hyporeflexia</td>
</tr>
<tr>
<td>(interpreted as ankle jerk and knee jerk less on left side)</td>
<td>Abdominal fullness sensation (bloating)</td>
</tr>
<tr>
<td>‘Feels ‘tightening’ along the sides of her abdomen and ‘tightening sensation in her upper abdomen’</td>
<td>Rovsing’s sign positive</td>
</tr>
<tr>
<td>‘Rovsing’s sign’, (sic)</td>
<td>Rovsing’s sign positive</td>
</tr>
<tr>
<td>Iron deficiency anemia</td>
<td>Iron deficiency</td>
</tr>
</tbody>
</table>

HIF, high-information clinical finding.
that future work identifying HIFs would still use an easily accessible knowledge resource such as an MDDSS KB as a starting point, but would evaluate the utility of the chosen set of HIFs iteratively in the disease and control populations.

**POTENTIAL LIMITATIONS OF THE WORK**

The physician reviewers were aware of the patient’s diagnosis during the record review and so their sensitivity to HIFs in the notes would be greater than it would have been were the diagnosis not known. However, as the major objective of this study was to determine, for patients known to have an HRD, whether HIFs were present in the clinical notes before the HRD had been established, it was desirable for the physician reviewers to have high sensitivity for detection of HIFs so as to be able to uncover the HIFs during this manual review. As we see this manual review as a necessary precursor to future work designed to evaluate automated extraction of HIFs, a very high sensitivity of the manual review is necessary to serve as the ‘gold standard’ for future automated systems.

It is possible that the physician reviewers determined that HIFs were present before the diagnosis had been established, when in fact the clinician had already considered the diagnosis. The criteria used to decide whether the diagnosis was established (diagnosis documented at future visit but not mentioned previously in note and not on problem list) seemed the best proxy available for this determination.

It is possible that using controls from the same dataset of 623 patients, even with different HRDs from the index patients, introduced bias that underestimated the difference from the index patients. For example, abdominal mass is an HIF in both ovarian and colon cancer. To the extent that a control patient with another HRD may be more likely to have this or some other HIF (than, say, a completely healthy patient), this reduces any difference that may be found in the comparison with the index patient. Because it is difficult to find a research data repository of records of completely healthy patients, we used our initial data cohort. This possible bias, which potentially underestimates the difference, is less concerning than a bias that might overestimate the difference found.

Because as many as a quarter of the records we reviewed contained HIFs in the clinical notes before documentation of the HRD, we feel that discovering these clues as early as possible could be of potential benefit to patients. Given the current progress with NLP tools, we are optimistic that automated extraction of these HIFs is a plausible goal in the foreseeable future, which may lead to a diagnostic decision support tool that could be integrated into the clinical workflow. Clinicians may not be receptive to diagnostic decision support or other alerts. It will be important to balance the thresholds at which a decision support tool generates an alert so as to minimize false positives, and consequently reduce alert fatigue. These thresholds will need to take into account the degree of risk of the diagnosis, the degree of confidence that the patient has the disease diagnosed, and the degree of certainty that the clinician has not already considered this diagnosis. In addition, the method in which the alert is presented will be an important consideration. However, even the best decision support tool will not fulfill its potential if it has no effect on clinician behavior. Fortunately, there are many studies, some from 20–30 years ago, that have demonstrated the ability of computer-based decision support systems to influence clinician behavior. In exploring the extent to which diagnostic decision support programs (QMR and liid) affected clinicians’ differential diagnoses, Friedman et al found a 6% increase in the correct diagnoses (according to a consensus of clinical experts) that appeared in the subjects’ differential diagnosis lists of diagnostically challenging cases after use of the program. The effect was greatest on medical students, intermediate on residents, and least on faculty. McDonald et al found that a computerized system that offered reminders to clinicians significantly increased the clinician response rate. In a prospective cohort study of house staff at a Mayo Clinic teaching hospital, 72% of residents surveyed responded that they ‘frequently or almost always’ considered novel alternative diagnoses after using DXplain, and 28% responded that they did so ‘occasionally.’

This preliminary work demonstrates that clinical clues can be present in medical records before the documentation of HRDs. This leaves us optimistic that, by collaborating with colleagues experienced in NLP, we may proceed toward the automated extraction of HIFs, which may allow integration of a diagnostic decision support tool into the clinical workflow. The use of standardized clinical vocabularies clearly will be helpful in this effort.

**Acknowledgments**

The authors thank Hang Lee, PhD, for statistical assistance.

**Funding**

Funding for this study was provided by a grant from the Partners-Siemens Research Council. The Partners-Siemens Research Council did not participate in the design of the study, in the collection, analysis or interpretation of the data, in the writing of the report or in the decision to submit the article for publication. The researchers were independent from the Partners-Siemens Research Council. This work was conducted with support from Harvard Catalyst, The Harvard Clinical and Translational Science Center (NIH Award #UL1 RR 025758 and financial contributions from Harvard University and its affiliated academic health care centers). The content is solely the responsibility of the authors and does not necessarily represent the official views of Harvard Catalyst, Harvard University and its affiliated academic health care centers, the National Center for Research Resources, or the National Institutes of Health.

**Competing interests**

MJF, EPH, GOB, RJK, KTf and HCC are employed by the Laboratory of Computer Science at Massachusetts General Hospital, the developers of the DXplain diagnostic decision support system. During the previous 3 year period, the Laboratory of Computer Science has received licensing royalties related to the DXplain system from approximately 40 hospitals and medical schools, and from Merck & Co, Inc and Epocrates, Inc.

**Ethics approval**

Partners Healthcare Human Research Committee.

**Provenance and peer review**

Not commissioned; externally peer reviewed.

**REFERENCES**


