

# Factors in Deploying Automated Tools for Clinical Abstraction and Coding

Mark L. Morsch, David S. Byrd, Daniel T. Heinze

*A-Life Medical, Inc, San Diego, CA, USA*

## Abstract

*In this paper, we explore human and technical factors involved in justifying, selecting and successfully deploying tools that automate clinical abstraction and coding. Throughout the US, healthcare providers are facing the challenge to improve processes and provide accountability in the areas of quality of care, billing practices, utilization of resources and information privacy. All of these initiatives, in one way or another rely on information that must often be abstracted from medical records. Applying information technology can enable an abstraction process that is accurate and cost-effective, producing results that are meaningful and accessible for a variety of purposes. However, the human factors are as important as the technical ones in successfully deploying automation into the healthcare information workflow. We review the current state-of-the-art in automated clinical abstraction and coding, discuss the criteria for selecting a solution, describe issues faced in deployment and present a case study of a medical billing organization that has deployed an automated coding solution.*

**Keywords:** *Abstracting, Automation, Information Management, Cost Control, Quality Assurance, Health Insurance Reimbursement, CPT Codes, ICD Codes, SNOMED, Natural Language Processing, Artificial Intelligence*

## Introduction

In the US, healthcare providers are collecting and managing an increasing amount of clinical data to satisfy the regulatory requirements related to the quality, cost and methods of patient care. Recent new laws and rule changes require more extensive monitoring of quality of care in hospitals, thereby motivating efforts to collect and report information related to specific treatments and outcomes. In medical billing, which specifies an encoding of certain clinical facts, both governmental and private payors in the US are tightening coding requirements for medical conditions to substantiate the medical necessity of many treatments. To monitor costs, hospital administrators are asking for systems that track the consumption of resources across a facility. Health Insurance Portability and Accountability Act

(HIPAA) guidelines mandate the use of standardized coding systems for billing purposes and define the rules for safeguarding patient health information. This trend is increasing the amount of information that must be captured and codified. Information technology can assist in making the abstraction and coding process more accurate, productive and secure, but maximizing these improvements requires an approach that best integrates the people, process and technology.

## Motivation

Fundamentally, abstracting and coding are necessary in the healthcare information workflow in order to extract the data elements from patient records required for administrative, regulatory and research purposes. Each of these applications may use different schemes for encoding, storing and reporting data. For example, for physician billing in the US, the International Classification of Diseases 9<sup>th</sup> Revision (ICD-9) [1] and Current Procedural Terminology (CPT) [2] codes are used; for quality reporting in hospitals, clinical “core measures” [3] as defined by the Joint Commission on Accreditation of Healthcare Organizations (JCAHO) are used; and for clinical record-keeping and research, SNOMED CT [4] may be used. Typically the original form of the patient record, whether it is paper, electronic or some combination of the two, will not represent the complete information from the record in a form consistent with some or all of the required encoding schemes. To bridge that gap, a human abstractor/coder will read and interpret the patient record, assigning the appropriate codes according to the guidelines of the particular encoding scheme. It is clear that these functions need to be performed with a high level of accuracy and consistency while maintaining audit controls and ensuring patient privacy. Information technology can be applied to help reach all of these goals in a cost-effective manner. However, IT solutions are more successfully deployed in organizations that have clear expectations, an understanding of current processes and the collective willingness and ability to bring about change.

The process of abstracting and coding varies greatly depending upon the organization and the specific purpose, but the following ten workflow functions are common to nearly all.

- Selection – Retrieval of patient records from a repository, often based on a particular disease or demographic profile
- Collating/Sorting – Matching of separate documents to form a complete record, organized into logical sequence
- Filtering Duplicates – Identification and removal of duplicate documents within a record or duplicate records within a batch.
- Distribution/Routing – Dissemination and transmission of patient records between parties involved in different workflow functions
- Abstracting/Coding – Identification and encoding into discrete data elements of information from the patient record
- Quality Assurance Review – Assessment and verification of the outcomes of the various workflow functions
- Data Entry – Manual transcription of information, typically the output of a function like abstracting/coding, into a system for subsequent processing and storage.
- Export/Interfacing – Establishing connections for data exchange between systems.
- Reporting – Creation of meaningful data representations for clinical or administrative purposes
- Exception Handling – Processes for handling situations that fall out the typical workflow, such as incomplete records, contradictory information or processing delays

Note that abstracting/coding, the step where the discrete codes or data elements are extracted from the records, is included here as one function of the overall workflow. Problems with any of the functions listed here can result in lost or corrupt information, introducing error into the final results. We will explore these further in the Selection Criteria section.

Probing into the workflow to focus on automation of the abstracting and coding function, we examine some characteristics of the task and describe the types of technology that best apply. We argue that the primary basis for justification of automation is in the complexity of the abstracting task itself. Medical abstracting/coding requires an understanding of medical vocabularies, anatomy, medications, disease processes, treatment protocols and experience with the various schemes for encoding this information. In the process of abstracting a record, human abstractors must make judgments based on their interpretation of the information in the record. In the Selection Criteria section, we present research demonstrating that the accuracy and consistency of these individual judgments drive overall quality and consistency in abstracting, and that individual differences in knowledge and experience can cause substantial variations in results. In the next sections we discuss how information technology can standardize these judgments and reduce errors.

## Selection Criteria

### Operational Expectations

Healthcare information technology decisions are accelerated by identifying how the proposed solution will advance strategic goals. However, management principles are only realized within an organization when goals are integrated with business efforts. Generally, the decision to deploy an automated tool for clinical abstraction and coding is made to eliminate the problematic bottlenecks associated with creating an end result. Traditional processes significantly impede human productivity due to paper based activities and the monotony that is involved in capturing and codifying clinical information. Yet operational expectations vary from organization to organization due to resource utilization and process protocol to conclude a compliant end result. Understanding the individual traditional processes involved in producing an end result will assist in ascertaining realistic expectations for automating clinical abstraction and coding. To do this the process analysis must start at the inception of the clinical document and followed through to the concluding result. Establishing metrics to traditional abstract/coding practices, prior to the implementation of the automated process will substantiate the automated solution. Post comparison analysis will identify the reduction of time for producing an end result, improved abstracting/coding productivity and consistency, and a reduction in overall associated cost.

### Workflow

Enterprise-wide workflow management systems have been deployed as part of electronic patient record (EPR) systems [5, 6]. Streamlining the workflow of the abstracting/coding process requires a thorough understanding of the individual functions and the environment within which they operate. Ideally, a computer-based process will automate or assist in each of the major functions. In general, workflow is enhanced when the number of electronic and paper touch points is minimized and tedious, manual processes are automated or eliminated. However, automation can be misapplied if the system does not fully meet all of the necessary requirements, particularly in the areas of quality assurance and exception handling. Manual processes can benefit from the ingenuity of people to handle special cases, and identifying those special cases is an important part of the requirements definition process. We present some of the key selection criteria and discuss their relationship with the ten workflow functions introduced earlier.

### *Number and Types of Systems to Integrate*

To deploy an IT solution for automated abstracting/coding, information must be exchanged between systems. This requires an incoming feed of demographics, to identify and index the records, and clinician documentation, to extract relevant data elements. The outgoing data feed, in its simplest form, is the encoded data elements. Electronic interfaces that integrate an

automated abstracting tool with other systems are desirable to reduce or eliminate data entry effort both incoming and outgoing. It is important to identify early in the specification and procurement cycle the points of integration between systems and the types of interfaces needed. Issues such as proprietary systems, shortage of resources or organizational resistance to change can impede an integration effort.

### ***Completeness of the Electronic Record***

Integration addresses access to the patient record, but a closely related issue is how much of the record is actually accessible in electronic form. An electronic environment addresses many of the failings of paper-based medical records [7], and abstracting/coding can leverage the investment in EPR systems. However, it is given that for most healthcare facilities at least part, and often the majority, of the patient record is still on paper. Different abstracting and coding tasks require different portions of the record. For example, coding for radiology physician billing may only require access to the radiologist report and certain demographic fields, while abstracting of data elements for JCAHO core measures requires access to a complete record including the discharge summary, history and physical examination notes, progress notes, medication orders and diagnostic test reports. Matching the availability of electronic information with the requirements of the abstracting task can minimize the role of paper in the workflow.

### ***Productivity of the Individual Coder/Abstractor***

Turning to the productivity of the individual user, advanced technology can be employed to reduce the burden of mundane and tedious tasks, such as record selection and sorting, and help the user quickly find the key information and make accurate judgments. XML interfaces can simplify the task of aligning demographic data fields, facilitating the automation of record assembly and collation. Natural Language Processing (NLP) technology can be used to extract information from transcribed physician notes and present the result to the user for verification. Consolidating the presentation of multiple parts of an electronic record into one user interface is another feature that will reduce time spent reading and searching the record. As we will discuss later on in the Case Study, IT products with these features have demonstrated productivity, quality and financial benefits.

### ***Automating Versus Assisting***

In introducing the ten workflow functions, we stated that problems with any of the steps can introduce error in the final results. Effective automation will improve productivity and reduce error. However, for tasks as complex as abstracting and coding, it is generally not a practical goal to fully automate every step of the workflow. Three functions in particular where computer assistance (or partial automation) is the best approach with current technologies are quality assurance review, exception handling and abstraction/coding. The common characteristic of each of these functions is the high level of interpretation and decision making required. Later on we will argue that for tasks such as abstracting/coding, humans and

machines working together complement each others strengths. This approach also maintains the expert oversight of the process, an essential feature in fulfilling healthcare compliance programs.

### ***Measuring Performance***

Finally, tools to report on and manage the process are critical. These tools can leverage the extensive database of information collected and stored during the automated abstracting or coding process. Productivity and quality reports, combined with workflow management tools, enable administrators to better manage resources and measure outcomes. This is an area where IT solutions can offer significant value, because with manual processes this data is often unavailable or too expensive and time-consuming to collect.

### ***Compliance***

Healthcare organizations are ethically and legally bound to maintain high standards for the quality of patient care, billing practices and patient privacy. Compliance programs ensure guidelines are followed by defining and documenting operational processes, training personnel that implement the process, implementing audits and developing performance metrics to measure results, and providing means for reporting and resolving issues [8]. IT solutions should advance the goals of an organization's compliance program.

Coding and abstracting are particularly important functions when considered in the context of a compliance program. Coding is typically performed as part of the billing process for a healthcare provider, and accurate coding is necessary to bill correctly. More generally, abstracting is performed to support a variety of important functions including billing, quality reporting, outcomes analysis and research. These activities supply facts that become the basis for conclusions such as the justifiability of treatment, success or failure of a procedure and disease trends.

Automation of clinical abstracting and coding, when deployed properly, has significant benefits to an organization's compliance program. A major benefit is the improvement in accuracy of the results, particularly for high volume activities that require a consistent application of guidelines, such as coding for physician billing. The two research studies presented later in the paper provide data to support this view. Anecdotally, it is accepted that there are significant differences in performance among human abstractors, particularly for those with different levels of experience and training. A structured workflow with results provided by NLP software is a more consistent process producing a more consistent result. Another benefit is the capability to monitor results and assess the level of conformance to guidelines utilizing reporting tools. This function provides a highly observable process, something not possible in a paper-based environment.

## Standards

For abstracted information to attain maximal value, it should conform to nomenclature standards that allow the sharing of data across platforms and users. Standards follow need and as would be expected, the older and more pervasive standards relate to coding for purposes of government records keeping and coding for billing. The various incarnations of ICD, including both diagnostic and procedure codes, are probably the most widely used. In the US, the American Medical Association's CPT is also ubiquitous for coding in the ambulatory care setting. Additionally, there are dozens of coding nomenclatures that either fill a particular niche or which were developed based on some organizations dissatisfaction with the otherwise available nomenclatures.

Most of the significant clinical nomenclatures have been widely reviewed and evaluated [9,10,11,12,13]. Some, notably SNOMED, have undergone significant evolutionary changes in order to meet the growing understanding of how nomenclatures can meet the needs for clinical information normalizing and sharing. The evolution of SNOMED has been in both extent of the nomenclature and more notably in the representation. Although it may not be a completely fair comparison, because there are other differences related to intended purpose, it is at least worthy of note that from ICD-9 to ICD-10, where essentially the same representational character was maintained, the ability to capture clinical information actually declined, whereas SNOMED made a far superior showing [14].

The choice of nomenclature has significant effects on the cost and effort for implementation and, more importantly, will lock an organization into a system with theoretical limits in terms of information representation capabilities. The following factors, at least, need to be considered when deploying automated tools for clinical abstraction and/or coding.

Will the system be used for reimbursement/billing coding and regulatory reporting only, or will it [also] be used for clinical abstraction with medical applications? If the system's use is restricted to billing and regulatory reporting purposes, the required nomenclature systems and the requisite storage requirements will be considerably simpler, at least for the present. This is because the nomenclatures for billing and reporting use, almost uniformly, pre-coordinated coding systems. A pre-coordinated coding system is one in which the various notions that make up a complex clinical concept are distilled to a single code. ICD, CPT and many others are pre-coordinated coding systems. A clinical notion may be composed of elements related to disease process, anatomical location, etiology, acuity etc., but all of these are captured by a coordinated description with a single code and no means to decompose the code or the description into its constituent concepts. Pre-coordinated codes make for simple storage and processing requirements, but they have drawbacks. Finely delineated codes result in combinatorial explosion, hence concepts in pre-coordinated systems will be either limited in

overall scope or will be individually overly broad with the effect that important aspects and distinctions in meaning are lost.

Until recently, most nomenclatures for abstraction also relied completely on pre-coordination. Advances in logic programming systems during the 1970's and 80's led to a surge in research and development of description logics in the 1990's to the present. Development was fueled in the late 1990's by the information representation and intercommunications needs of web-based business applications. Description logic (DL) allows a compositional approach to nomenclature. Complex concepts can be composed or coordinated by linking together simple concepts using relationships (roles). For example, in SNOMED CT, a simple procedure concept, say a surgical incision can be coordinated with an anatomical site via the relationship (role) of associated topography, with a scalpel via the uses equipment relationship and so on. Because concepts can be coordinated after they have been observed in clinical usage (post-coordination), coding can be as detailed and finely specified as needed. Because description logics can support comparison, categorization and semantic distance measures, codes that are coordinated at different sites and for different needs have a basis for comparison and information sharing.

The implementation of a description logic clinical abstraction system comes with its own costs and limitations. First, the storage of ad hoc post-coordinated concepts is a task that few commercial data base systems can handle. Relational data base (RDB) systems have become pervasive over the past twenty years, but DL is more naturally represented in an object oriented data base (OODB). Relatively speaking, RDBs are fast, simple, efficient and manageable, but when they are configured to hold DL representations, they lose all of these desirable qualities. Unless commercial systems that can efficiently handle DL storage and operations are widely adopted, DL nomenclatures like SNOMED CT will be restricted to the use of only their pre-coordinated concepts and little benefit will be realized. The shortcut that seems to be most promising is to use XML to represent DL concepts. XML fields can be stored in RDBs and handled separately or by the growing capabilities of XML processors that are being built into RDBs. The approach can be extended from the concept level to the clinical document level by employing an emerging XML-based standard such as the Clinical Document Architecture [15].

DLs do have a self-imposed limitation, however. It is in the number and types of roles. For good but limiting reasons that are beyond the scope of this paper to discuss, SNOMED CT can be thought of as a language with all the nouns and adjectives you want but with only a handful of verbs and prepositions and without the adverb "not". The relationships (roles) of a DL can be compared to the verbs, prepositions and conjunctions of human language. For both theoretical and practical reasons, the number and type of relationships (roles) in DLs must be severely limited. This means that some concepts will be either difficult or impractical to express in a DL. The objective of the SNOMED committee is to choose the allowable relationships in

such a manner as to strike an acceptable balance between expressiveness and computability.

For the decision maker, the factors are: 1) know what kinds of concepts you need to abstract or code; 2) understand the large variety of nomenclatures that are available; 3) consider what the needs for interoperability and data/information sharing between departments, sites and organizations will be; 4) explore the representational capabilities of both the automated abstracting systems and the data base backbone that are under consideration; 5) choose solutions that balance flexibility with cost, because data representations may evolve on a shorter cycle than the replacement cycle of your information systems.

### Comprehension and Accuracy

The discussion of standards for medical coding and abstracting illustrates the fact that the level of comprehension required in order to perform coding and abstraction is largely dictated by the target nomenclature and also the selected codes within the nomenclature. In order to tag individual words, it is only necessary that the automated system disambiguate the senses of polysemous words. As the complexity of the nomenclature increases, the demands on the language processing system increase. It is relatively easy to find all the occurrences of the term “smoker” in a set of documents. It is harder to find all the discussions of patients who are smokers when the documents contain information about family members who smoke, second-hand smoke etc. It is even harder to find patients with a “history of smoking” when patients who stopped smoking a year or more ago are excluded, e.g. “... had a 20 pack month history but stopped in July 2003.”; report dated May 2004. The system must know that “20 pack month history” refers to smoking history, must consider the date of cessation and must compare that with the date of the examination.

Measuring accuracy in complex coding and abstracting tasks is not easy and no one should presume that there is a gold standard against which to test. Objective tests that exclude biases and unintended external input must be devised, and statistical methods should be employed for evaluating test results. Much can be learned from the literature on test design and on statistical quality control. Here we will make just a few observations and present a pair of research studies.

One way to restrict the effects of bias and unintended input is to make the test size so large and the test time so short that it would be impossible to produce anything other than uncompromised automated coding or abstraction results. The problem, however, is in how to evaluate the results. In the first research study, we present a large volume test with an evaluation of overall results against expected results and also a statistical sampling against manual abstraction. Another technique is to use a smaller test set with strict controls over the test procedure and with multiple participants. We exemplify this method in research study 2. From these research studies, we note that the relative strengths and weakness of both human and automated coders and abstractors are such that with correct

system design and implementation, a result can be obtained that is greater than the sum of the parts.

### Research Study 1 [16]

In order to test automated abstraction for a task in support of pharmaceutical research, a test case consisting of 53,656 medical notes from across the range of ambulatory and acute care clinical settings and specialties at four major university medical centers and one private medical center was selected. Three disease profiles with specific abstraction targets were defined. These were acute myocardial infarction, asthma and gallbladder disease. Each profile required the mining of demographic information, primary diseases, co-morbidities, medications, medical and/or surgical interventions, and outcomes. The clinical abstraction specifications for the acute myocardial infarction profile are summarized in Table 1.

<ol style="list-style-type: none"><li>a. Identify survival rate during hospitalization.</li><li>b. Group by presence of risk factors, e.g. family history, hyperlipidemia, diabetes mellitus, cigarette smoking, prior myocardial infarction.</li><li>c. Identify presence of co-morbidities, e.g. valvular disease, chronic obstructive pulmonary disease.</li><li>d. Group by presenting symptoms, e.g. chest pain, dyspnea, arm/leg pain, jaw pain.</li><li>e. Group by location of infarction: anterior, anteriolateral, inferior, right ventricular.</li><li>f. Group by type of infarction: transmural vs. non-transmural.</li><li>g. Group by duration of hospitalization.</li><li>h. Identify 5 most common types of arrhythmias present during hospitalization.</li><li>i. Group by use of thrombolytics.</li><li>j. Group by use of aspirin after hospital presentation; where possible, identify time interval between hospital presentation and administration of aspirin.</li><li>k. Identify patients undergoing acute cardiac catheterization versus later.</li><li>l. Group catheterized patients by major anatomic abnormalities: left main, LAD, left circumflex disease: and significant one vessel, two vessel or three or more vessel disease.</li><li>m. Group by discharge medications: aspirin, beta-blockers, anticoagulation.</li></ol>
---

Table 1: Acute MI Text Mining Requirements

The tests were designed so that crosschecks were performed to validate results. For example, the severity of an MI or an asthma attack was to be assessed both by the stated assessment of severity as given by the clinician and also by the type and number of treatments administered and the need for follow-up.

A summary of results is presented in Table 2. Manual sampling of the results validated a ~99% accuracy level. To achieve this high accuracy, certain specificity constraints had been relaxed as follows. Several items requested in the profiles were determined, even before the test, to be beyond the ability of the abstraction engine as it existed at the time of the test. In some cases this was because accurate determination of the fact often required the unification of information given across multiple

documents – e.g. the determination of whether a gallbladder removal was emergency or scheduled. In other cases this was because the information was not reliably reported in the clinical documentation at hand – e.g. how soon after the onset of an MI the patient was administered aspirin. Finally, some information would have required a development effort beyond the scope allowed for in the test – e.g. determining not just that a laparoscopic technique was used for a gallbladder removal, but also which of the four specific techniques defined by CPT was used. In total, of the 39 major categories of information specified in the three profiles, there were two urgency-related items, one technique-related item and one timing-related item that could be extracted with an accepted level of accuracy only by generalizing the profile requirements for those items.

Number of documents	53,656
Number of sources	5
Query profiles and number / percent of encounters identified for each.	Acute myocardial infarction – 854 / 1.6%
	Acute exacerbation of asthma – 1695 / 3.2%
	Gallbladder disease – 372 / 0.7%
Number of queries	39
Accuracy	~99%
Processing platform	550 MHz Pentium III
Processing time	~10 sec / document

Table 2: Research Study 1; Summary of Results

These results indicate that with certain constraints, automated extraction with a very high level of accuracy can be achieved, even for fairly complex target concepts and nomenclatures. Further, it is possible for the automated abstraction system to know the limits of its capabilities so as to enable a synergistic integration with human abstractors.

**Research Study 2 [17]**

To illustrate the use of a closely controlled test with a more limited test set size, consider a test of evaluation and management coding performance by an automated coding system as compared to both production coders and expert consultants in emergency medicine coding. In order to avoid bias from the study sponsors and also to keep the human coders from being biased by knowing the source and purpose of the study, the authors contracted an outside market research firm to select, contact and handle the transactions with the participating coders. The marketing firm was presented with a master list of coders and was asked to select a representative sample from among individual experts, premium billing companies and standard billing companies. Six human coders were selected to participate in the study, and one coding expert was selected to act as an auditor to blindly collate and review the results from the six coders and LifeCode, an NLP processor that assigns ICD-9 and CPT codes [18].

One hundred transcribed emergency medicine charts were selected for the study and were sanitized by removing all personal identifying information. Half of these charts were selected to be representative of cases that are most common in emergency medicine and half were selected at random. To help eliminate potentially confounding factors in the experiment, all coders were required to agree in writing to the following guidelines:

1. All 100 records must be coded in a single session of no more than 7 hours, inclusive of breaks.
2. All charts are to be coded. If for some reason a chart cannot be completely coded, coders are to move on to the next chart and not return to the uncompleted chart.
3. Charts must be coded in the order presented, and once a chart has been coded it cannot be reexamined for review or changes.
4. The coding session shall start at the coder's normal business day start time but can be on a weekend if needed.
5. Coding shall be conducted in an environment that is free from interruptions and distractions.
6. All coding shall be done according to the Health Care Financing Administration (HCFA) 1995 Documentation Guidelines for Evaluation and Management (E&M) Services. A copy and summary of these guidelines is given to the coders for use as a reference.
7. ICD-9 codes are to include both E and V codes, and CPT codes are to include E&M codes and modifiers.
8. Coding is to include the correct sequencing of both ICD-9 and CPT codes.
9. Paper-based ICD-9 and CPT manuals, the attached summary of the 1995 HCFA guidelines, and a medical dictionary are the only reference materials to be used. The summary of the HCFA guidelines should be reviewed before starting the study.
10. Consultation with other individuals should neither be sought nor accepted.

All participants, including the auditor, coded each and every chart. Each chart was also processed in LifeCode to produce both an E&M code and the associated diagnoses and procedures codes.

The study participants recorded the CPT E&M level of service codes (99281 - 99285) for emergency medicine for all 100 charts. Analysis of inter-coder agreement was done using the Kappa statistic that provides a conservative comparison of how well one coder agrees with another taking into account chance agreement [19]. Kappa is calculated by dividing the difference between observed and random agreement by the maximum possible difference between observed and random agreement:

	A	B	C	D	E	F	G	H	Consensus Agreement	Consensus Kappa	-2	-1	+1	+2	RVU
<b>A</b> (Expert)	*	0.35	0.24	0.31	0.34	0.43	0.51	0.34	<i>0.73</i>	0.57	3	19	4	0	2.00
<b>B</b> (LifeCode)	<i>0.58</i>	*	0.24	0.30	0.36	0.33	0.36	0.28	<i>0.71</i>	0.51	0	13	12	0	2.13
<b>C</b> (Standard Billing)	<i>0.51</i>	<i>0.55</i>	*	0.28	0.57	0.37	0.27	0.30	<i>0.72</i>	0.57	3	18	4	0	2.17
<b>D</b> (Expert)	<i>0.51</i>	<i>0.52</i>	<i>0.49</i>	*	0.30	0.45	0.40	0.37	<i>0.68</i>	0.46	3	12	11	0	2.14
<b>E</b> (Expert)	<i>0.55</i>	<i>0.63</i>	<i>0.72</i>	<i>0.50</i>	*	0.44	0.38	0.30	<i>0.78</i>	0.63	0	7	11	0	2.18
<b>F</b> (Premium Billing)	<i>0.58</i>	<i>0.53</i>	<i>0.55</i>	<i>0.57</i>	<i>0.65</i>	*	0.33	0.34	<i>0.69</i>	0.54	0	2	18	1	2.36
<b>G</b> (Standard Billing)	<i>0.67</i>	<i>0.58</i>	<i>0.54</i>	<i>0.58</i>	<i>0.59</i>	<i>0.53</i>	*	0.38	<i>0.71</i>	0.54	3	18	2	0	1.95
<b>H</b> (Expert/Auditor)	<i>0.51</i>	<i>0.49</i>	<i>0.52</i>	<i>0.55</i>	<i>0.48</i>	<i>0.53</i>	<i>0.54</i>	*	<i>0.59</i>	0.42	3	21	8	0	1.99

**Table 3:** Research Study 2; A-H Inter-rater agreement (*italics*) and Inter-coder Kappa (standard); Agreement with the consensus; Kappa with the consensus; Number of charts deviating -2/-1/+1/+2 from consensus; Average RVU for EM levels.

$$K = \frac{P_0 - P_r}{1 - P_r} \quad (1)$$

where  $P_0$  is the observed agreement,  $P_r$  is the random agreement, and 1 is the maximum observed agreement. A Kappa value less than 0.40 indicates poor agreement. Kappa between 0.40 and 0.70 indicates fair to good agreement. Kappa above 0.70 indicates strong agreement. Since raw agreement rates (i.e., how often coder A agreed exactly with coder B in the aggregate) are widely used, results are also summarized using both exact agreement and Kappa. Each coder was also compared to the study group consensus.

Inter-coder Kappa and agreement results are presented in Table 3. In columns A-H, Table 3 reports the pair-wise comparisons between each coder including LifeCode and the auditor. Coders A, D and E are nationally recognized experts in emergency medicine coding. Coder F is a coder assigned from a premium billing company that is nationally recognized for quality. Coders C and G are coders assigned from standard, competent billing companies that are considered typical of the industry. Coder B is LifeCode. Coder H is the auditor who has an expert status comparable to coders A, D and E.

The left-most matrix in Table 3 is laid out as two half-tables: on the lower left are the agreement rates between the coders; on the upper right are the corresponding Kappas.

The Consensus columns in Table 3 show the score of each coder against the consensus opinion on each chart. Because there were eight participants coding each chart, if there was a tie then the auditor's code was removed from consideration in order to break the tie. The columns -2 to +2 report on the number of charts per coder that deviated by the indicated magnitude and direction from the consensus. In other words, if coder A scored a chart with 99283 but the consensus scored the same chart with 99281 then A would receive a +2 rating for that chart because A coded two levels higher than the consensus.

The RVU column represents the average relative value unit per the E&M codes of each coder. RVUs are used for comparison instead of the average E&M level because the value scale is not

linear. Here the E&M level to RVU correspondence is per the December 1999 Federal Register [20]. Coders B, C, D and E all lie less than one-half standard deviations from mean RVU of 2.12, whereas A, G and H are one or more standard deviations below the mean and F is more than two standard deviations above the mean.

The differences in inter-rater Kappas as compared to individual rater versus the consensus are significant. The mean inter-rater Kappa is 0.38 (poor) with a standard deviation of only 0.08, whereas the mean rater versus consensus Kappa is 0.53 (fair to good) with a standard deviation of 0.06. The mean inter-rater Kappa plus one standard deviation (0.46) is still less than the mean rater versus the consensus minus one standard deviation (0.47). Because no individual had a strong Kappa ( $> 0.70$ ) as compared to the consensus, we would hesitate to raise the consensus to a "gold standard". This hesitancy seems further justified by the observation that it is possible to be very close to the consensus Kappa but significantly above (coder F) or below (coder G) the mean RVU level.

Expert study participants show the least agreement between themselves. In terms of consensus agreement, LifeCode is about average amongst participants. With respect to the consensus, LifeCode produced the most balanced distribution of codes, whereas expert coder A had the most prominent downward skew and Premium Billing Company F had the most prominent upward skew.

If the consensus level for each chart is used as a gold standard, then the codes for most charts should cluster around the consensus and be evenly balanced one level above and below the consensus with a few outliers [21]. However, this is not the case. The typical pattern, as shown in Table 3, is significant deviation from the consensus, with clusters one level above or one level below rather than being balanced on both sides of the consensus.

Among production coders in hospitals and billing companies, the most common check on coding accuracy is the level of agreement between a coder and some other coder who is accepted as an expert or standard. If the level of agreement is calculated between coders within the same organization who

have been trained in the same practices and are held accountable to the same internal coding standards, the level of agreement can be quite high. Chao [22] noted a very high E&M agreement level for nurses trained in a similar methodology in a primary care environment. In order to achieve a high agreement level, methodology training was required and direct observations were obtained by the nurses themselves (coding was not done from the physician's notes).

Anecdotally, we note that billing companies generally claimed agreement rates of 95 percent to 98 percent. Lloyd and Layman [23] report similar agreement rates. However, these are situations where the supervisor is reviewing the coder's work – not double-blind studies. In the declining number of cases where billing companies send a sample of their charts to another company (one selected for similarity of philosophy and practice), much lower agreement rates are achieved. Rates on the order of 80 percent to 85 percent have been reported to the authors, but double blind outside audits of billing company codes versus the auditor show agreement in only the 50 percent range [24].

### ***Coding and Abstracting with Humans and Machines***

Can humans and machines work together on coding and abstracting tasks in a manner that produces job satisfaction for the human and measurable performance and financial improvements for the overall organization? We believe so, but success depends on an understanding of the relative strengths and weakness of human and machine abstractors and an implementation that minimizes the weaknesses and emphasizes the strengths of each.

Subjectivity in medical records coding results from the wide variety of medical conditions reported, the language used by the caregiver, the various interpretations of coding guidelines, and the complexity of the documentation guidelines. Human coders are required to extract a large number of details and distill these details into the few codes that accurately describe the patient's visit.

Differences of opinion may not necessarily mean that one person is right and the other is wrong. A coder's perspective or work environment often influences the way charts are interpreted. Because coding guidelines are general rules for measuring and categorizing the work of clinicians, the specific application of the guidelines is subject to human interpretation. Published guidelines contain, at best, one or two prototypical examples of each code. From these guidelines alone there is no way to address every possible real-world instance. Abstractors and coders must apply their best judgment and this is the seed of defensible disagreement. These disagreements will exist whether the abstractors and coders are human or machine. There are, however, definable errors and these can be typed and categorized.

There are broadly two categories of errors that coders make in decision-making tasks, performance errors and systematic errors [25,26,27,28,29]. Performance errors consist of mistakes and

slips. In the realm of medical coding, mistakes involve misreading words in the source document or missing details altogether, resulting in a misunderstanding of document content. An example of a mistake would be failing to see negation or failing to pull details together from across the document. Slips involve failing to carry out the intended coding assignment, for instance the transposition of two digits in a code.

Systematic errors, on the other hand, consist of knowledge- and rule-based errors. Knowledge-based errors occur because the coder lacks the medical knowledge to grasp the situation at hand – often due to inadequate training and experience in the field. Rule-based errors involve a misapplication of rules. The coder may understand the document and correctly assess the medical case it describes, but incorrectly assign the code due to confusion or lack of knowledge about the coding standards.

Performance errors, like "mistakes" or "slips," are rare in computerized coding systems, but systematic errors can occur due to errors in the system's knowledge bases or training examples. Also, the ability to create generalization and handle novel expressions or idiosyncratic and idiomatic terminology is less than for humans.

To successfully integrate human and machine abstractors and coders, the following guidelines should be considered and accounted for in the workflow and system design and implementation. 1) Make sure that the automated system has the ability to assess its own performance and invoke human assistance when needed; 2) Where human intervention is needed, the automated system should provide an adequate explanation of why the intervention is need; 3) The automated system should provide easy access to as much information as possible; 4) If the automated system makes systematic errors, there should be reasonable remedies whereby they can either be corrected or trapped; 5) It should be possible to determine why the machine makes any particular judgments. With these guidelines in mind, implement a system that emphasizes machine performance with human supervision and, as needed, aid when the machine reaches the limits of its knowledge. Particulars of the people and process issues involved are discussed next.

### **Maintenance and Support**

For an IT solution to be successful and deliver a return on investment, it must continue to work over time. This requires an accurate assessment of the amount and type of ongoing maintenance and support. Healthcare requires robust solutions that often must function in a 24/7 environment. Automated tools for abstracting and coding are no exception to this, particularly to meet the demands of real-time clinical decision-making. Healthcare facilities may need to develop additional in-house expertise to support these tools as they become more pervasive in the clinical setting.

Another factor is the level of customization that is provided by automated abstracting or coding tools. These tools have medical

domain knowledge built-in, so as guidelines and rules change, the tools must change. Coding guidelines for ICD-9 and CPT are updated annually. Additional quality measures are under development by JCAHO and will be introduced in the next few years. Facilities will, over time, change the set of data elements that need to be abstracted. The healthcare IT planner should understand how the various tools support these changes and associated time, effort and cost to implement.

## **Deployment Factors**

### **People Issues**

Introducing an automated tool to the medical coder and/or clinical abstractor has both technical and social factors. On the technical side, the tool's ease of use and dependability are critical. The user interface should be both flexible, to fit into different workflows, and efficient, to keep up with the experienced user. A well-designed tool will allow the end user to quickly review and edit the results of the automated process. However, an automation tool can be potentially threatening to the medical coder and/or clinical abstractor. The high accuracy of the delivered result and the productivity gains that are delivered by the tool can be perceived as justification for staff reductions. In every situation, product acceptance by staff should be considered early on in the specification and procurement cycle to avoid deployment difficulties.

Our experience has shown that healthcare organizations will fall into one of three types in handling the reduced labor requirements that result from automation. The first type has an existing personnel shortage and is incurring additional costs of overtime. In this case, automation allows the organization to maintain existing staff but likely reduce or eliminate overtime. The second type seeks to redeploy staff to address new demands for abstracting/coding resources. The third type will reduce staff, particularly in the lower-skill areas of data entry and administration assistance. Whatever the situation, the organization must be motivated to change by management. Resistance to this type of change is understandable, but staff should appreciate that eliminating the most tedious and mundane portions of the work will result in a more stimulating and rewarding work environment.

### **Process Issues**

Process issues encountered when implementing an automated abstraction and coding tool are as unique as the technology itself. Healthcare regulatory reporting, in most cases, is derived from multiple document sources from a consortium of healthcare providers by specified case types. Physician medical billing for reimbursement is obtained by understanding the documented patient encounter and translating the encounter into billable codes. In either situation there are two primary process issues when moving from manual abstracting and coding to utilizing the automated abstracting and coding tool.

First and foremost is the lack of readily available electronic documentation that is needed to automate the functions of the clinical abstraction and coding process. Typically when implementing an information system or computer application the data needed for processing is widely accessible. The healthcare industry has had a heavy reliance on hand written documentation. Due to this, if proper electronic documentation is not secured prior to the implementation of the automated tool, the deployment will be delayed.

Secondly, and perhaps more importantly, is the efficient utilization and acceptance of the technology by the end users. Deployment of an automated tool to the medical abstractor/coder offers a significant improvement in productivity. However, these benefits are realized only with a computer-literate staff that is willing to adopt new methods of doing work. A change in approach for the abstractor/coder is required in going from assigning results to reviewing results. Initial and follow-up training and regular monitoring of productivity metrics help ensure that the maximum benefit is realized. Additionally, introducing an automated tool into a specialized job task that has never experienced computerization can bring about additional ancillary process issues. These issues are encountered due to job tasks that have become translucent or seen as non-existent tasks that assist in obtaining the end result. These tasks may include the multitude of ways and the manner in which data is gathered, displayed, routed and stored.

### **Operational Metrics**

Successful deployment of an automated process for medical coding and clinical abstracting requires that the decision maker be able to identify financial and compliance gains. Consistency of the manner in which an automated process abstracts information permits a better means to manage and project financial outcomes. Deployment of such automation has a tremendous impact on overall operating cost by reducing payroll and cutting cost associated with paper related activities. The productivity gain experienced by automatically abstracting data and eliminating paper related activities is potentially very large, and the associated infrastructure assures a compliant and secure environment in which data is gathered, displayed, routed and stored. The electronic abstraction of data also creates a permanent audit trail of how the data was derived, thus reducing liability risks associated with governmental and private payor guidelines. The case study in the next section illustrates specific metrics for productivity, turn-around time, cost reduction and quality improvement.

## **Case Study**

### **Overview**

A forty five member physician group operating in the US southeast is generating approximately thirty five thousand transcribed physician notes per month. The physician group's billing staff consists of three full time data entry employees, and

three and half full-time coders. Coder productivity is averaging two hundred seventy-five physician notes per day.

### Issues

Staffing levels are deemed appropriate for the ratio of physicians to patient encounters. However, the physician group is exceeding forty hours of coder overtime per week and has a chronic backlog in administrative processing. This backlog is exacerbated by an inefficient and error-prone manual process of collating the physician notes with patient demographics. Also, manual data entry is required to enter the assigned ICD-9 and CPT codes into a billing system.

### Results

Significant benefits were experienced by the organization within a six month time period after deploying an automated tool for clinical abstraction and coding [30]. Table 4 summarizes the performances metrics for pre- and post-deployment. For workflow benefits, electronic interfaces allowed data entry staff to be reassigned and data alignment algorithms automated the collation of physician notes with patient demographics. Individual coder productivity increased by more than two-and-a-half times, eliminating weekly overtime and reducing coding staff requirements. The chronic backlog was eliminated, shaving twenty days from average charge entry days (the difference between date of service and the date the ICD-9 and CPT codes for the service were entered into the billing system), twenty days from the average number of days between the date of service and the posted reimbursement date and sixteen days from the average account receivable days.

Performance Metric	Pre-Deployment	Post Deployment
Charge Entry Days	30	10
Date of Service to Posted Reimbursement	50	30
Average Account Receivable Days	79	63
Data Entry Staff	3	0.5
Coding Staff	3.5	2
Coder Productivity Per Day	275	702
Overtime Hours Per Week	40	0

Table 4: Abstracting/Coding Process Analysis

### Conclusion

In conclusion, we highlight some commercial applications that provide a guidepost for parties interested in gauging the maturity of current solutions. As of mid 2004, there are at least three vendors in the US providing automated medical coding products that utilize NLP technology for coding radiology and emergency medicine. These products are employed at more than 200 facilities, and that number is increasing. Several major Health

Information Systems vendors have current initiatives in data warehousing, and one vendor has competed a beta test with a solution for automated abstraction of quality measures in support of JCAHO requirements. These developments offer strong evidence that we have moved out of the initial research and development phase. Commercially proven products are available now, and more will be developed to meet the growing demand.

### References

- [1] Puckett CD. 2004 Annual Physician Version: The Educational Annotation of ICD-9-CM. Fifth Edition. Reno, NV: Channel Publishing; 2003.
- [2] Current Procedural Terminology: CPT 2004. Fourth Edition. Chicago, IL: American Medical Association; 2003.
- [3] Joint Commission on Accreditation of Healthcare Organizations. Specification Manual for National Implementation of Hospital Core Measures Version 2.0. Available at: <http://www.jcaho.org/pms/core+measures/information+on+fi+nal+specifications.htm>. Accessed July 21, 2004.
- [4] SNOMED International. What is SNOMED CT. Available at: [http://www.snomed.org/snomedct/what\\_is.html](http://www.snomed.org/snomedct/what_is.html). Accessed July 21, 2004.
- [5] Murray M. Strategies for the Successful Implementation of Workflow Systems within Healthcare: A Cross Case Comparison. Proceedings of the 36th Annual Hawaii International Conference on System Sciences; 2003 Jan 6-9. p. 166-175.
- [6] Mahoney ME. Document Imaging and Workflow Technologies in Healthcare Today. Journal of American Health Information Management Association 1997 April;68(4):28-36.
- [7] Safran C, Goldberg, H. Electronic patient records and the impact of the Internet. International Journal of Medical Informatics 2000;60:77-88.
- [8] Ortquist S, Vacca S. Is Your Compliance Program Measuring Up: Guidelines Offer Practical Advice to Judge Effectiveness. Journal of American Health Information Management Association 2004 April;75(4):62-64.
- [9] Rose JS, Fisch BJ, Hogan WR, Levy B, Marshall P, Thomas DR, Kirkley D. Common Medical Terminology Comes of Age, Part One: Standard Language Improves Healthcare Quality. Journal of Healthcare Information Management 2001 Fall;15(3):307-318.
- [10] Rose JS, Fisch BJ, Hogan WR, Levy B, Marshall P, Thomas DR, Kirkley D. Common Medical Terminology Comes of Age, Part Two: Current Code and Terminology Sets – Strengths and Weaknesses. Journal of Healthcare Information Management 2001 Fall;15(3):319-330.

- [11] Rose JS, Kirkley D. Healthcare Computer Applications and The problem of language: A Brief Review. *Informatics Review* [serial online] 2000 Dec 15. Available from: <http://www.informatics-review.com/thoughts/cmt.html>. Accessed July 23, 2004.
- [12] Chute CG, Cohn SP, Campbell JR. A Framework for Comprehensive Health Terminology Systems in the United States: Development Guidelines, Criteria for Selection and Public Policy Implications. *Journal of the American Medical Informatics Association* 1998 Nov 1;5(6):503 - 510.
- [13] Blair JS. An Overview of Healthcare Information Standards. Computer-based Patient Record Institute electronic publication. Available from: <http://www.hipaonet.com/cpri.htm>. Accessed July 23, 2004.
- [14] Chute CG, Cohn SP, Campbell KE, Oliver DE, Campbell JR. The content coverage of clinical classifications. For The Computer-Based Patient Record Institute's Work Group on Codes & Structures. *Journal of the American Medical Informatics Association* 1996 May;3(3):224-233.
- [15] Clinical Document Architecture (CDA). Draft. Ann Arbor, MI: Health Level Seven, Inc.; December 2003.
- [16] Heinze DT, Morsch ML, Holbrook J. Mining Free-Text Medical Records. In: Bakken S editor. *Proceedings – AMIA Annual Symposium*; 2001 Nov 3-7; Washington DC; 2001. p. 254-8.
- [17] Morris WC, Heinze DT, Warner Jr HR, Primack A, Morsch AE, Sheffer RE, Jennings MA, Morsch ML, Jimmink M. Assessing the accuracy of an automated coding system in emergency medicine. In: Overhage JM editor. *Proceedings – AMIA Annual Symposium*; 2000 Nov 13-18; Los Angeles CA; 2000. p. 595-9.
- [18] Heinze DT, Morsch ML, Sheffer RE, Jimmink M, Jennings MA, Morris WC, Morsch AE. LifeCode: A Deployed Application for Automated Medical Coding. *AI Magazine* 2001 Summer;22(2):76-88.
- [19] Fleiss JL. *Statistical Methods for Rates and Proportions*. 2nd ed. New York: John Wiley & Sons; 1981.
- [20] Medicare Program; Revisions to Payment Policies Under the Physician Fee Schedule for Calendar Year 2000; Final Rule. *Federal Register* 1999 Nov 2. Available from: <http://frwebgate6.access.gpo.gov/cgi-bin/waisgate.cgi?WAISdocID=606291115666+1+0+0&WASAction=retrieve>. Accessed July 23, 2004.
- [21] Hripcsak G, Kuperman GJ, Friedman C, Heitjan DF. A Reliability Study for Evaluating Information Extraction from Radiology Reports. *Journal of the American Medical Informatics Association* 1999 Mar/Apr;6(2):143-150
- [22] Chao J, Gillanders WG, Flocke SA, Goodwin MA, Kikano GE, Strange KC. Billing for Physician Services: A Comparison of Actual Billing with CPT Codes Assigned by Direct Observation. *The Journal of Family Practice* 1998 July;47(1):28-32.
- [23] Lloyd SS, Layman E. The Effects of Automated Encoders on Coding Accuracy and Coding Speed. *Topics in Health Information Management* 1997 Feb;17(3):72-79.
- [24] Private audits performed for customers of A-Life Medical, Inc.; 1999.
- [25] Norman DA. Categorization of action slips. *Psychological Review* 1981; 88:1-15.
- [26] Norman DA. *The psychology of everyday things*. New York: Harper & Row; 1988.
- [27] Reason JT. Lapses of attention. In Parasuraman R, Davies R editors. *Stress and fatigue in human performance*. Chichester UK: Wiley; 1984.
- [28] Reason JT. *Human error*. New York: Cambridge University Press; 1990.
- [29] Wickens CD. *Engineering psychology and human performance*. New York: HarperCollins; 1992.
- [30] Florida Radiology Associates. Alamonte Springs, FL; 2004.

#### Address for correspondence

Mark L. Morsch, Vice President NLP/Software Engineering, A-Life Medical, Inc, 6055 Lusk Boulevard, Suite 200, San Diego, CA 92121, USA, [mmorsch@alifemedical.com](mailto:mmorsch@alifemedical.com)