A Modern Optical Character Recognition System in a Real World Clinical Setting: Some Accuracy and Feasibility Observations

Paul G. Biondich, M.D., J. Marc Overhage, M.D., Ph.D., Paul R. Dexter, M.D., Stephen M. Downs, M.D., Larry Lemmon, Clement J. McDonald, M.D.
Regenstrief Institute for Health Care and Indiana University School of Medicine, Indianapolis, IN

Abstract
Advances in optical character recognition (OCR) software and computer hardware have stimulated a reevaluation of the technology and its ability to capture structured clinical data from preexisting paper forms. In our pilot evaluation, we measured the accuracy and feasibility of capturing vitals data from a pediatric encounter form that has been in use for over twenty years. We found that the software had a digit recognition rate of 92.4% (95% confidence interval: 91.6 to 93.2) overall. More importantly, this system was approximately three times as fast as our existing method of data entry. These preliminary results suggest that with further refinements in the approach and additional development, we may be able to incorporate OCR as another method for capturing structured clinical data.

Introduction

Capturing structured clinical information is the first and most challenging step on the road towards developing an electronic medical record (EMR). In recent years, we have focused on the development of computer workstations for the capture of physicians' orders, clinical notes, and other data [1]. With the help of extensive user feedback, physician electronic order entry has been successful throughout the hospital and many clinics. We have also had some success with physician entry of clinical notes via this same workstation [2].

However, work styles, keyboard skills, workflow, and clinical content vary widely among physicians and specialties, and conversion process from paper to workstations remains slow. As a result, pen and paper remains an often used method of clinical note and data recording within our hospital system. Despite its well known legibility and logistic problems, paper maintains a strong hold on health providers. It is familiar, lightweight, flexible, and fast [3]. Today, we can include images of these documents in an EMR via document scanning. Furthermore, computerized interpretation of printed text or optical character recognition (OCR) technology offers the potential to structure this clinical information.

Given this background, we asked whether integrated OCR solutions might provide a good data capture solution that physicians would use in multiple clinical settings. Some have had successes with this technology [4,5,6]. Indeed, we explored OCR in the mid and late 70's when we recognized blood pressures and other clinical measurements off of paper encounter forms [7]. At that time the technology had substantial limitations. It required special pens or pencils, and the forms had to be preprinted using special inks. Further, the scanners of that time could not deliver a digital image for medical record storage. We also faced difficult logistic problems that came with physical delivery of the encounter forms from the clinics to a central reading station. This meant extra copying and forms that never reached the input station.

Recent advances in hardware and OCR technology have eliminated many of these problems. OCR no longer requires special pens or forms preprinted with special inks. Currently, the technology can read from forms printed on demand using black and white laser printers. The completed documents can be scanned at the point of care, and sent digitally to a centralized OCR reading station avoiding the logistic problems of copying and transportation. Furthermore, the scanned document can be saved as an image within the medical record. Given these advances, we decided to again test the feasibility and the error rates of OCR data capture off of paper using actual data in real world clinical settings. For this pilot study, the data consisted of hand written numbers recorded on a slightly modified version of our standard on-demand printed encounter forms.

Background:

The Regenstrief Medical Records System (RMRS) collects appointment information and provides the clinics with customized paper encounter forms [8]. Vital signs and other numeric observations are handwritten into fields displayed within a column along the left side of each form. Once the caregivers complete these encounter forms, they are photocopied and sent through campus mail to the Regenstrief Institute. Trained data entry specialists interpret handwritten values and key these results into a data capture program within the RMRS. This program then sends these results as an HL7 message to the RMRS data repository. In designing our pilot, we sought to eliminate the data entry step but otherwise retain the existing workflow.

Methods

We obtained approval for the study from the Indiana University Institutional Review Board which also serves as the IRB for Wishard Hospital.

Data Collection:

We conducted the study in Wishard Memorial Hospital's Pediatric Outpatient Clinic, located in downtown Indianapolis. Encounter forms were generated by the RMRS and completed by the clinical and support staff as they have been for over twenty years. Staff in the clinic fed the completed forms into a Digital Sender 8100C enterprise scanning device (Hewlett Packard, Palo Alto, California, http://www.hp.com) which was placed in the registration area of the clinic, along with a set of instructions. The Digital Sender accepts a batch of the completed encounter forms and then creates a multi-page tagged image format file (TIFF) containing the digitized output at 300 dots-per-inch resolution. These files are then emailed to a server in the Regenstrief Institute through a connection to our hospital's network located behind a firewall.

We used Teleforms Elite 7.1™ (Vista, California, http://www.cardiff.com), an integrated OCR software recognition engine, to process the handwritten numbers on each encounter form. The software reads both computer-printed and handwritten alphanumeric characters contained within recognition zones that are defined through creation of form specific templates. A copy of this software was installed onto our server computer using the Windows XP operating system, and Microsoft Outlook XP mail client. This server receives email from the Digital Sender and an interface included within Teleforms automatically retrieves the TIFF from Outlook's mailbox for processing.

Adaptation of Encounter Forms for OCR:

The RMRS generates patient specific encounter forms on the basis of such variables as the scheduled clinic, the patient's age, immunization history, etc. The existing forms had to be modified slightly for the OCR reader, as the actual space allotted for written entry was too small and provided no cues for proper spacing (Figure 1a). We initially increased the space allotted to the numeric input field and used "comb" style fields to help cue providers about appropriate sizing and spacing of characters (Figure 1b). In pilot testing it became clear that even more cues were required to specify the position of the decimal point and to add two part values such as pounds and ounces. The final version of the form used for our pilot study provided three blanks for the integer portion of the value, and a two blank extension to specify decimal position (Figure 1c).

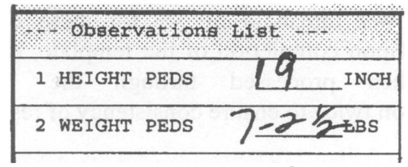

Figure 1a: An example of the original appearance of form fields.

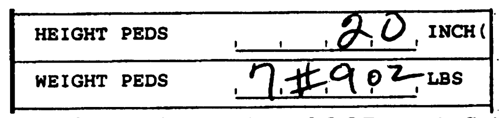

Figure 1b: A pilot version of OCR-ready fields.

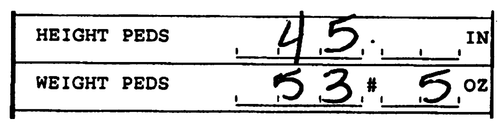

Figure 1c: Final version with explicitly denoted units.

Data Sources:

One hundred fifty forms were sent through the Digital Sender to our OCR server by pediatric clinic support staff over a six day period. Fifteen or more numeric observations are potentially recorded on these forms, but within this collection, three fields were completed for most visits: height (in inches), weight (in pounds and ounces), and head circumference (in centimeters). Nurses always enter the weight and height information and physicians measure and enter the patient head circumferences. Eight nurses and nineteen physicians completed forms during this six day period. Nurses were informed of the trial of the scanning system and received brief instruction; the physicians did not.

Data Evaluation:

A template of the encounter form was created using the "Designer" application within the Teleform package. We applied several preprocessing algorithms in order to remove the pre-printed comb fields which we added to guide input in each of the six fields of interest on this template. Each form has an upper and lower confidence threshold value. When a character is analyzed by the software, it is assigned a confidence value which corresponds to the OCR engine's certainty of the digit identity. If the digit's confidence value exceeds the form's upper threshold, the software accepts it as "correct". If the confidence lies between the two thresholds, the software makes its best guess and flags the digit for review. For confidence values below the lower threshold, the software does not attempt a guess, and enters a placeholder character for correction by the data verifier. We chose an upper threshold of 95% and a lower threshold of 5% for these encounter forms. Upon completion of the template, the forms were then processed through the "Reader" application twice to ensure consistency of results.

To evaluate the software's accuracy, we tallied the total counts of digits and numeric values on all submitted forms and classified each as nurse or physician-entered. Each form was then manually reviewed in the "Verifier" application by one of the authors (PB) to establish a gold standard for correct readings.

Each digit's result was classified based on whether it fell into the high (>95%) confidence range and required no verification, fell below the low (<5%) confidence range and had to be manually entered, or whether the recognized digit required review (5 – 95%). All of these results were entered into a spreadsheet along with some descriptive comments of the OCR software's successes and failures.

To evaluate the usability and speed of the computerized verification process, we gave printed copies of the 150 encounter forms to our most experienced data entry clerk for traditional manual entry. After approximately two hours of instruction and orientation, we then asked that same clerk to process these forms through the Verifier application after processed a third time by the OCR engine. Both methods of data acquisition were timed for comparison.

Results

The 150 forms in this study contained a total of 982 digits and 564 different numerical values. These 982 digits represented 805 distinct digits entered by nurses and 177 digits entered by physicians.

Overall, the software was able to recognize a digit with greater than 95% confidence 58% of the time (Table 1). Digits written by nurses were assigned a higher average confidence than physicians. Approximately 42% of all values required some form of review (either low or medium range).

Table 1: Confidence Assignments (% of totals)

	High (>95%)	Medium (5-95%)	Low (<5%)
Nurses (n=805)	482 (59.9%)	315 (39.1%)	8 (0.9%)
Physicians (n=177)	86 (48.6%)	78 (44.1%)	13 (7.3%)
Totals (n=982)	568 (57.8%)	393 (40.1%)	21 (2.1%)

The vast majority of all written digits were correctly recognized by the software (92.4%, 95% confidence interval (CI): 91.6 to 93.2). There were no errors for the 453 digits assigned confidence values of greater than 95% (Table 2).

Table 2: Accuracies at Various Confidence Levels (% Correct)

	Non Review (>95%)	Review (5 – 95%)	Overall
Nurses	482/482 (100%)	275/315 (87.3%)	757/805 (94.0%)
Physicians	86/86 (100%)	64/78 (82.0%)	150/177 (84.7%)
Total	568/568 (100%)	339/393 (86.3%)	907/982 (92.4%)

Digits entered by nurses were recognized more accurately (94.0%, 95% CI: 93.2 to 94.8) than those entered by physicians (84.7%, 95% CI: 82.0 to 87.4) (Table 2). When looking at numeric values, 499 of 564 were read correctly (88.5%, 95% CI: 87.2 to 89.8). These results were consistent in both sets of data.

139 of the 150 forms (92.7%) had at least one digit that would have required verification in a real world deployment using our confidence thresholds. 75 individual recognition errors were made by the software but 43% were due to cross outs, entry of fractional values, and failure to follow directions (Figure 2):

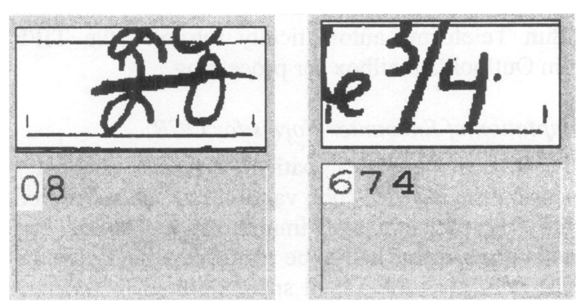

Figure 2: Examples of the types of user errors leading to recognition errors.

The data entry clerk took approximately 36 minutes to manually enter the information from paper copies of the encounter forms. This time included a manual review of the typed results. The OCR system, on the other hand, allowed her to complete the task in approximately 12 minutes.

Clinic staff appeared to easily incorporate the Digital Sender into their workflow. Although some TIFF images arrived skewed and even inverted, the software corrected their orientation in all cases making them adequate for evaluation and archival. They also had an excellent resolution and were artifact free.

In addition, the staff preferred the modified version of the encounter form, even though this form had slightly less space available for the remainder of their note.

Discussion

Given the vagaries of human writing, OCR's inherent error rate will likely always be greater than zero. Manual review of a subset of interpreted results will consequently be necessary. In our pilot study, we found a 8% overall error rate in numeric recognition. Evaluation of those 75 recognition errors showed that physician recorded digits were more error prone (27/177 or 15%) than those entered by nursing (48/805 or 6%). In this study, we neither informed the physicians that the forms would be read by OCR software nor provided training. As a result, recognition errors within these forms were introduced. Physicians would attempt to write fractions out and exceed the boundaries of the recognition zones (Figure 3).

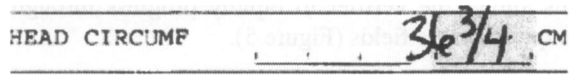

Figure 3: Example of user introduced recognition error. The recognition zone is highlighted.

It seems likely that these and other types of errors could be decreased through feedback and training and by dedicating somewhat more space to the recording fields.

Although the error rate is significant, it is made much more tolerable given the robustness of the automated system's verification process. In our current system, we capture data through trained entry clerks who key the patient medical record number, then select the kind of encounter form and enter the values they read on the left side of the encounter form. These typed values must be printed and finally verified against the original reports. With the Teleform package however, this process is simplified and faster.

The verification application shows a major part of the form and highlights the variable that needs review along the top portion of the screen (Figure 4). In the lower portion, the computer's interpretation of the digits is placed in close proximity to an enlarged picture of the field of interest as written by the staff. The digits are color coded to highlight those that need review.

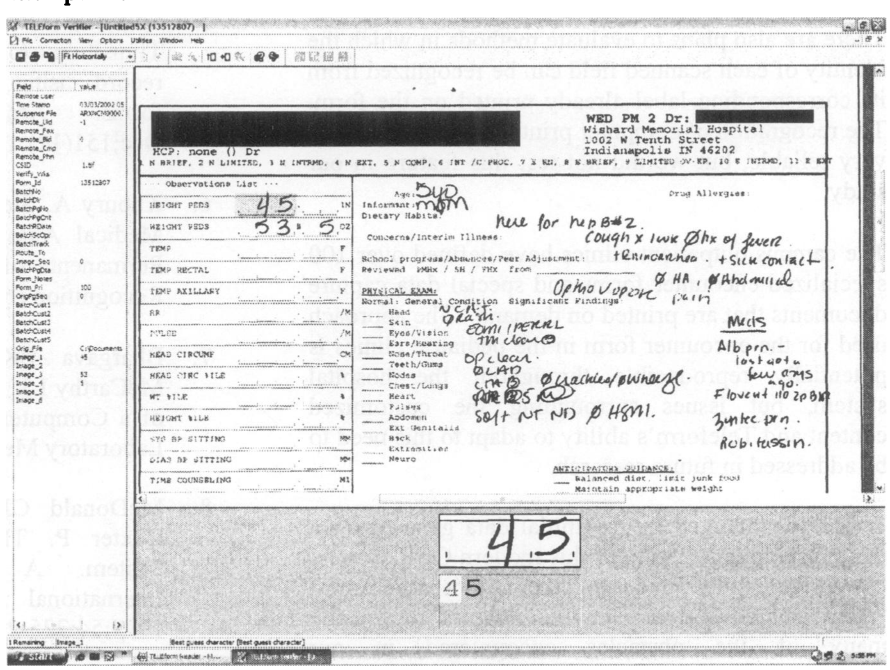

Figure 4: An example of a screen from the verifier application within Teleform. The screen layout provides an easy way to quickly review the recognition engine's questionable responses and make quick corrections if necessary.

With our configuration, a press of the tab key indicates acceptance of the software's best guess. This allows the verifier to rapidly progress through the questionable fields (Figure 5).

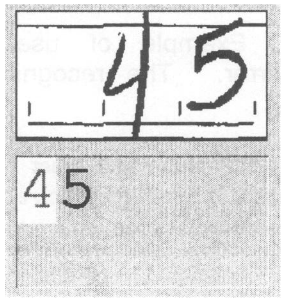

Figure 5: An example of a field within the verifier. In this case, you can just tab through this field to advance.

In this pilot trial, we demonstrated that this verification package, along with the software's inherent recognition rate, allowed the data entry clerk to verify and correct misrecognized values three times as fast as our current fastest method of data entry. This significant improvement in human time requirements is particularly impressive given the other benefits the system provides, most notably the digital copy that can be stored within the data repository.

More work is necessary to make this process ready for production. To tie the software's output to the particular patient's electronic record, we need to define other fields on the encounter form template to recognize computer-printed patient identifiers. There are also plans to evaluate methods in which the identity of each scanned field can be recognized from its corresponding label already printed on the form. The recognition of computer printed text is reportedly very reliable, but we did not test this feature in our study.

The campus outpatient clinics have defined over 100 specialized encounter forms and special data capture documents that are printed on demand. The approach used for the encounter form in the pediatric clinics is potentially reproducible throughout the hospital system, but issues surrounding the customized content and Teleform's ability to adapt to this need to be addressed in future research.

We are encouraged by the initial data gleaned from the pilot implementation of this modern OCR system. Use of the Digital Sender together with a more robust "form aware" software engine appears to remedy many of the obstacles encountered in our implementation 20 years ago. Future studies and research will be dedicated to continued testing and working towards automating the data flow back into the RMRS through a HL7 interface.

Acknowledgements: This work was performed at the Regenstrief Institute for Health Care in Indianapolis, Indiana and was supported in part by a grant from the National Library of Medicine (T15 LM-7117-05).

References

1. McDonald CJ, Tierney WM. The Medical Gopher--a microcomputer system to help find, organize and decide about patient data. West J Med 1986; 145(6):823-829.

2. Overhage JM, Perkins SM, Tierney WM, McDonald CJ. Controlled Trial of Direct Physician Order Entry: Effects on Physician Time Utilization in Ambulatory Primary Care Internal Medicine Practices. JAMIA 2001; Vol. 8:No. 4.

3. Sellen AJ, Harper RH. The Myth of the Paperless Office. MIT Press 2002.

4. Shiffman RN, Brandt CA, Hoffman M, Wiig W, Fernandes LA. SEURAT: Scanned entry of structured data for a pediatric health maintenance record. In: Masys DR, ed. Proceedings of the Annual Meeting of the American Medical Informatics Association; Nashville (TN); 1997; on CD-ROM.

5. Shiffman RN, Brandt CA, Freeman BG. Transition to a computer-based ambulatory record using scannable structured encounter forms. Arch Pediatr Adolesc Med 1997 Dec;151(12):1247-53

6. Khoury A, Siemon C, Mills G, Kalata M. The Medical Automated Record System of Kaiser Permanente of Ohio. Proceedings of the CPR Recognition Symposium 1997:21-100.

7. Bhargava BK, McDonald CJ, Rivera HP, McCarthy LJ. Development and Implementation of a Computerized Clinical Laboratory System. Laboratory Medicine 1976; 7(12):28-37.

8. McDonald CJ, Overhage JM, Tierney WM, Dexter P. The Regenstrief Medical Record System. A quarter century experience. International Journal of Medical Informatics 1999;54:225-253