Extracting Medical Information from Paper COVID-19 Assessment Forms

Colin G. White-Dzuro*[1] Jacob D. Schultz*[1] Cheng Ye[2] Joseph R. Coco[2] Janet M. Myers[3] Claude Shackelford[3] S. Trent Rosenbloom[1,2] Daniel Fabbri[2]

[1] Vanderbilt University School of Medicine, Nashville, Tennessee, United States
[2] Department of Biomedical Informatics, Vanderbilt University Medical Center, Nashville, Tennessee, United States
[3] Department of Medicine, Vanderbilt University Medical Center, Nashville, Tennessee, United States

Address for correspondence Colin G. White-Dzuro, BA, Vanderbilt University School of Medicine, Nashville, TN 37232, United States (e-mail: colin.g.white-dzuro@vanderbilt.edu).

Appl Clin Inform 2021;12:170–178.

Abstract

Objective This study examines the validity of optical mark recognition, a novel user interface, and crowdsourced data validation to rapidly digitize and extract data from paper COVID-19 assessment forms at a large medical center.

Methods An optical mark recognition/optical character recognition (OMR/OCR) system was developed to identify fields that were selected on 2,814 paper assessment forms, each with 141 fields which were used to assess potential COVID-19 infections. A novel user interface (UI) displayed mirrored forms showing the scanned assessment forms with OMR results superimposed on the left and an editable web form on the right to improve ease of data validation. Crowdsourced participants validated the results of the OMR system. Overall error rate and time taken to validate were calculated. A subset of forms was validated by multiple participants to calculate agreement between participants.

Results The OMR/OCR tools correctly extracted data from scanned forms fields with an average accuracy of 70% and median accuracy of 78% when the OMR/OCR results were compared with the results from crowd validation. Scanned forms were crowd-validated at a mean rate of 157 seconds per document and a volume of approximately 108 documents per day. A randomly selected subset of documents was reviewed by multiple participants, producing an interobserver agreement of 97% for documents when narrative-text fields were included and 98% when only Boolean and multiple-choice fields were considered.

Conclusion Due to the COVID-19 pandemic, it may be challenging for health care workers wearing personal protective equipment to interact with electronic health records. The combination of OMR/OCR technology, a novel UI, and crowdsourcing data-validation processes allowed for the efficient extraction of a large volume of paper medical documents produced during the COVID-19 pandemic.

Keywords

- COVID-19
- data processing
- optical mark recognition
- optical character recognition
- data creation and storage
- crowdsourcing
- medical form extraction

* *Authors contributed equally to this study.*

received
September 10, 2020
accepted after revision
December 25, 2020

Georg Thieme Verlag KG,
Rüdigerstraße 14,
70469 Stuttgart, Germany

DOI https://doi.org/10.1055/s-0041-1723024.
ISSN 1869-0327.

Background and Significance

The SARS-CoV-2 (COVID-19) pandemic has disrupted standard health care practice and forced health systems to change how they deliver care. For example, nonemergent surgeries are being postponed, telehealth use for routine outpatient care has surged, and clinics have employed texting services to enable patients to wait in their cars until their appointment.[1,2] Given the large volume of health care workers infected with COVID-19 worldwide (up to 15% of all cases in some countries[3]), health care workers evaluating patients with COVID-19 symptoms have also adapted to the current situation by routinely using personal protective equipment (PPE) during each visit.[4] Full PPE can be cumbersome to wear and limit the health care worker's ability to interact with electronic health record (EHR) systems. Assessing potential cases of COVID-19 and documenting both clinical findings and patient histories into an EHR while wearing PPE can slow and complicate the intake assessment processes. In this setting, paper-based clinical documentation may be preferred over and more efficient than computer-based documentation.[5] However, paper-based documentation may reduce the amount of computable data available for reuse, such as for disease-specific registries and decision support.[6] Additionally, medicine is becoming increasingly data driven and ensuring that pertinent health information ends up in a patient's EHR is crucial for physicians to employ holistic and preventative care on both an individual and population-wide level.[7]

Thankfully, there are methods by which paper forms can be rapidly transformed into data and stored in a patient's EHR. This relies on optical mark recognition and optical character recognition (OMR/OCR) software, which use image processing and computer algorithms to convert paper-based notations into digital data. These software have been used and evaluated for decades.[8–14] Past studies have demonstrated recognition of handwritten characters with >90% accuracy.[5,10,11,15,16] Technology has since advanced to allow for much more accurate and cost-effective usage of OMR/OCR.[17–19] In addition, research demonstrates that the overall human time requirement for data collection is decreased using OMR/OCR, while provider satisfaction for paper forms structured for digital analysis is equal or higher.[10,12,16] Early studies determined that nearly 30% of forms examined by OMR/OCR methods had to be verified by human personnel before being added to formal data systems such as EHRs.[5]

Objective

This study aimed to implement paper intake forms and OMR/OCR of paper-based COVID-19 assessment forms in novel combination with crowdsourced human validation to allow for safe, efficient clinical data collection that could be quickly converted to a digital form for storage and analysis.

Methods

Setting

All assessments recorded in this study were performed at Vanderbilt University Medical Center (VUMC), a large and tertiary medical center in Nashville, TN and its affiliated clinics which include 137 ambulatory care clinics. The EHR used by VUMC and in this study was the Epic electronic health platform (Epic Systems Corporation, Verona, WI). Patients were included if they received a nasopharyngeal swab for SARS-CoV-2 infection.

Population

A total of 10,369 COVID-19 intake forms were completed and scanned between March 16 and April 27, 2020. The assessment form evolved rapidly within the first 2 weeks, with the third and final iteration designed to better accommodate the OMR/OCR technology while maintaining clinical relevance (**Supplementary Material 1–3** [available in the online version]). This most recent assessment form accounted for 4,400 of the 10,369 scanned documents. Among those, 2,814 have been analyzed by one of the five crowd participants, with 483 documents analyzed by 2 crowd participants (the remaining 1,586 are queued to be reviewed). Forms analyzed by two crowd participants were included to analyze the agreement between observers (►**Fig. 1**). The number of analyzed assessment forms is currently limited by the number of crowd participants, which should increase with additional recruitment processes. The crowd participants continue to annotate these documents at the time of submission.

Assessment Form Development

A team that included authors, informaticists, operational leadership, and front-line health care workers from COVID-19 assessment sites developed a paper-based assessment form for clinical documentation as part of COVID-19 assessments. This form was designed to maximize efficiency of use by health care workers and support OMR/OCR technology. Clinical findings presented on the form were sorted into categories including history of present illness (e.g., fever and cough), review of systems, physical exam (e.g., blood pressure and temperature), and orders/ diagnoses/plan. To support subsequent optical recognition, each finding on the assessment form was aligned with a checkbox that could be marked rather than circled. Because there was significant variability in the circles/boxes made on forms by health care workers, this alignment helped eliminate the errors in data stemming from the sizes of circles/boxes. In addition, the form minimized narrative-text components to improve in-room usability; sections where narrative text was permitted were transformed into text boxes to normalize the area in which health care workers could write. To increase compatibility with OMR software, handwritten fields were further minimized in later forms to limit the variability of responses and improve the ability of the OMR software to read the marks. A total of 141 findings were included on the form.

After institutional review board and operational approval, the assessment forms were implemented in clinical assessment sites to evaluate patients with potential COVID-19 infection. Upon being completed in clinic, the assessment forms were scanned into EHR by using standard workflows for other paper-based clinical documents.

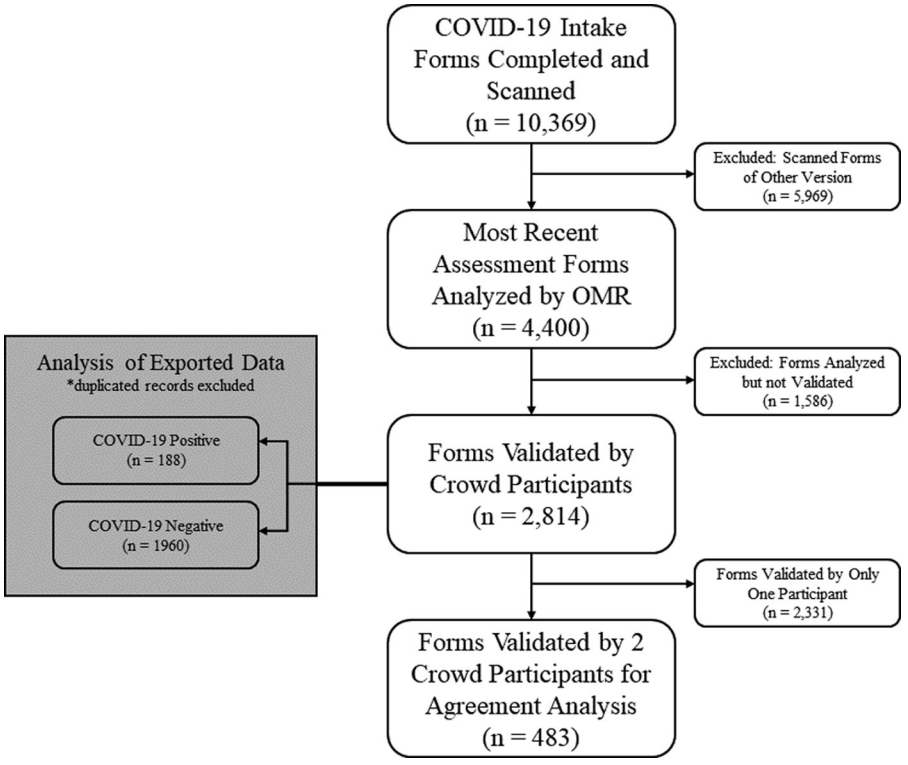

Fig. 1 Flow chart of exclusion criteria and subgroup analyses.

Optical Recognition and Validation

The Optical Mark Recognition/Optical Character Recognition Methods Worked as Follows

- First, the scanned paper assessment forms were aligned to a standard template by using OpenCV.[20,21] The alignment was accomplished by computing a homography matrix using a scanned form and the template, and then rotating the scanned form to until aligned.
- Second, a pixel map was created to track the (x, y)-coordinates of each field on the intake form. Each form version has its own pixel map.
- Third, OpenCV's HoughCircles function is used to detect circles (and other marks) on the scanned forms. The function returns a list of circles within the form, where the circles must have a radius between a prespecified minimum and maximum.
- Fourth, the detected circles' and marks' (x, y)-coordinates are compared with the pixel map to determine if a form field was selected or not, respectively. Due to the imperfect nature of circles on forms, we used multiple heuristics to assign detected circles to a field: (1) assigning a circle to the field with the largest overlap and (2) if there was no overlap (possibly due to imperfect alignment), choosing the field with the pixel-map coordinate that is closest to the circle or mark.

The OMR/OCR output was then reviewed for accuracy by a crowdsourcing mechanism.[22] Ideally, the OMR/OCR would be 100% accurate, but due to human handwriting under time constraints, perfect extraction is not possible. Instead, the goal of the OMR/OCR system is to accurately extract a large proportion of the fields quickly, thus allowing the crowd workers to quickly validate results and clean up mistakes. From our experience, a worker can more quickly validate the OMR/OCR results than manually reviewing a raw scanned form and inputting the results themselves.

The crowdsourcing system provides a scalable framework for data validation. New participants can easily be added to the participant pool over time. Moreover, the system can control the number of duplicate reviews per form. For data quality and access control, a subset of assessment forms was reviewed by multiple crowdsourcing participants. Participant agreement was measured by calculating a ratio of the number of exact matches between participants and the total number of entries. Any text input was converted to lowercase. Initially, workers were requested to transcribe handwritten comments; however, clinic leadership indicated that digitizing the narrative comments was not necessary for COVID-19 treatment and not performed by the OMR/OCR in this study, so the directive to transcribe those data was removed, further motivating the decrease in handwritten fields in later forms.

Timing data for start and completion of each of these tasks were stored by the crowdsource platform. However, these participants were not actively monitored; thus, it cannot be known if the participants took breaks. Therefore, the interquartile range was used to filter out timing data, which were greater than 401.5 seconds. For this project, the crowd consisted of medical students and physicians, and strict security controls limited the degree to which participants could access the data.

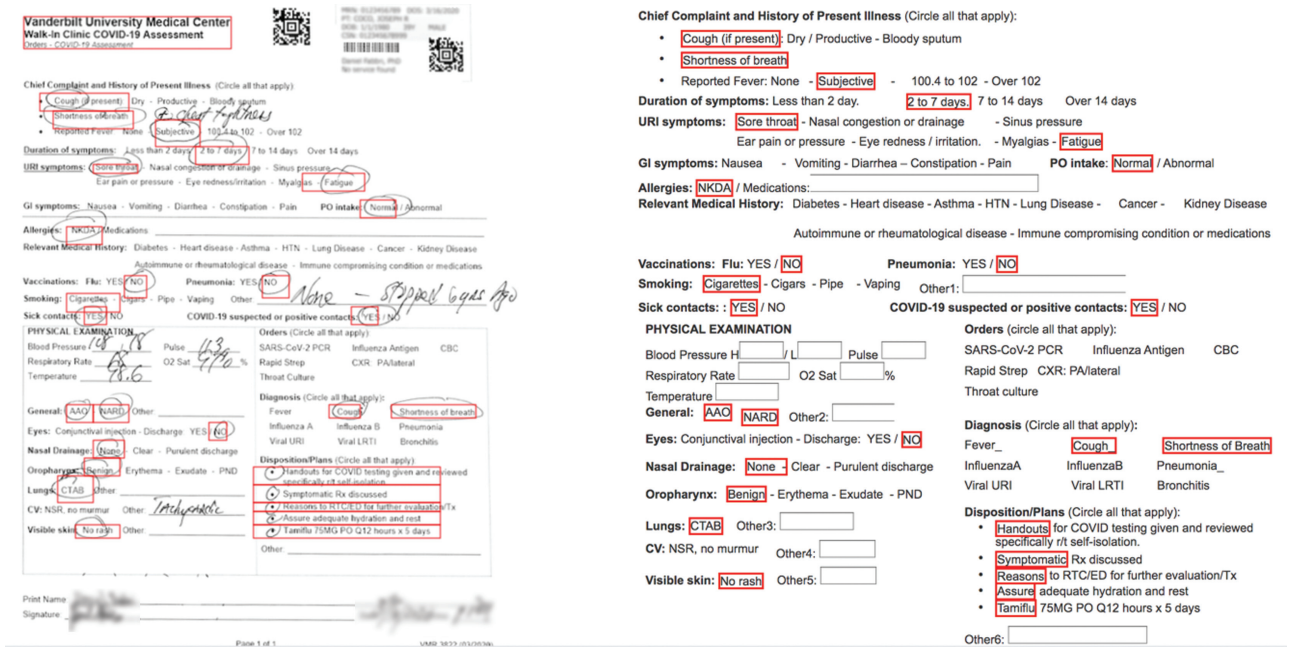

Fig. 2 Crowdsourcing user interface: (left) scanned page of early form with physician circling with optical character recognition results overlaid. (right) HTML form with selected fields marked.

The user interface (UI) of the crowdsourcing system had two main components. On the left half of the screen, the system displayed the scanned document with the OMR/OCR results overlaid. On the right half, an editable web form was displayed, which mirrored the structure of the paper document. All markings overlaid on the left were selected on the form on the right (►Fig. 2). This design allowed for quick confirmation of the OMR results and a "what you see is what you get" confirmation that the data are captured accurately in the structured web form input.

To allow for rapid user input, both the left and right screens could be clicked to add or remove a selection. For example, by clicking on the scanned form's "sore throat" element on the left, the field would be bounded by a red box and the corresponding field on the right screen would also be immediately boxed. This paired view enabled much quicker reviews compared with preliminary implementations that (1) did not have a mirrored form on the right or (2) had a form on the right, but the fields were not aligned in the same manner (i.e., all the fields were contained in a single column, making the crowd participant scroll down the page).

The crowd-validated data were then uploaded to a REDCap (Research Electronic Data Capture) database.[23,24] Study data were collected and managed by using REDCap electronic data capture tools hosted at VUMC. REDCap is a secure, web-based software platform designed to support data capture for research studies, providing (1) an intuitive interface for validated data capture; (2) audit trails for tracking data manipulation and export procedures; (3) automated export procedures for seamless data downloads to common statistical packages; and (4) procedures for data integration and interoperability with external sources. Data were only used for research purposes and were not exported to providers or the EHR for clinical use.

To demonstrate the efficacy of the OMR/OCR and crowdsourcing process, raw data were exported from the REDCap database and analyzed. All patients younger than 18 years of age were excluded from the study. Vital sign values determined to be nonphysiologic (systolic blood pressure <80 or >225 mmHg, diastolic blood pressure <40 or >140 mmHg, heart rate <40 or >140, respiratory rate <8 or >30, oxygen saturation <80 or >100, and temperature <96 or >104°F) were excluded. SARS-CoV-2 polymerase chain reaction (PCR) testing results were obtained via institutional sources. Patients were sorted into categories based on status of SARS-CoV-2 testing (detected, not detected, not tested). For those patients tested multiple times, the most recent test result was used.

Statistical Analysis

Data failed normality testing, so nonparametric testing was used throughout. Mann–Whitney tests were used for continuous data and Fisher's exact tests were used to test categorical values. A p-value <0.05 was used to determine significance. All statistical calculations were generated with GraphPad Prism version 8.0.0 (www.graphpad.com).

Results

►Fig. 3 represents the number of COVID-19 intake forms of various types that were crowd validated and uploaded to our analysis server. The crowdsourcing platform is currently bottlenecked by the number of crowd participants who have processed an average of 134 documents per day (►Fig. 4).

The mean time took a crowd participant to process a document was 157 seconds. The lower quartile was 108 seconds, median 156 seconds, and upper quartile 229 seconds. Crowd participants were previously taking

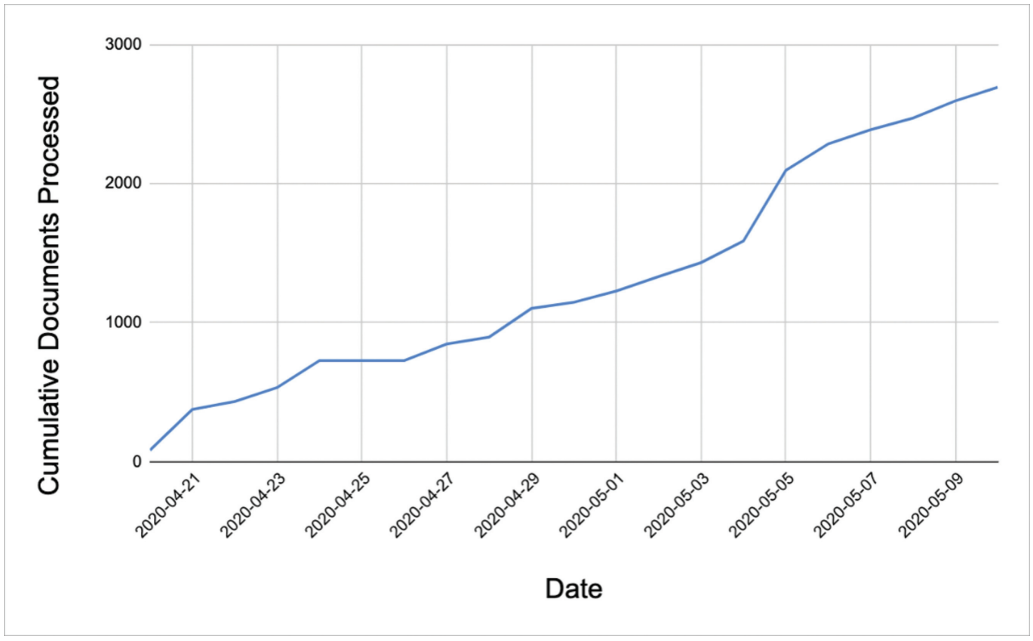

Fig. 3 Cumulative volume of COVID-19 intake forms over time.

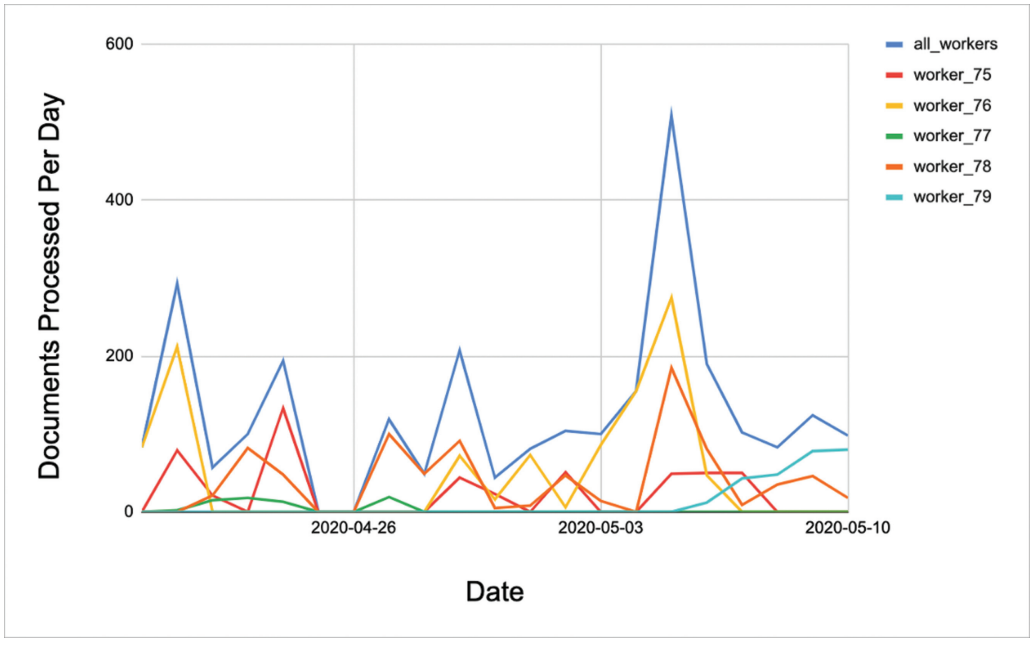

Fig. 4 Crowd-sourced documents processed by date and worker.

an average of 188 seconds to complete annotations; however, after removing the handwriting directive, they took an average of 156 seconds.

Each document contained 141 findings, 112 of which were presented as multiple choice style questions (i.e., cough is presented as "if present" with multiple choice modifiers including "dry," "productive," and "bloody sputum") and 29 as narrative-text, where the user could enter either a numerical value (e.g., vitals) or free text to elaborate on symptomatology. Of the 483 documents analyzed by multiple workers, the interparticipant agreement when including narrative text was found to be 97.6%. However, when comparing only multiple choice fields, the interparticipant agreement rose to 98.7%. Disagreement was spread among questions, which would imply human error or ambiguity as the cause.

The accuracy of the OMR/OCR tool was calculated to measure the effective reduction in manual data entry. ►Table 1 shows the average accuracy of the OMR/OCR tool by comparing its result with the result of each crowd participant. Overall, the OMR/OCR tool achieved an average accuracy of 70% and a median accuracy of 78%. The OMR/OCR tool provides a good start point for extracting information from scanned intake forms, while the crowd workers only need to correct a small proportion of results.

Table 1 Optical character recognition accuracy broken down by crowd participant

Crowd participant	Average accuracy (standard deviation)	Median accuracy
1	0.68 (0.26)	0.77
2	0.72 (0.24)	0.79
3	0.77 (0.18)	0.81
4	0.66 (0.27)	0.76
5	0.71 (0.25)	0.79

Table 2 Representative data collected by optical mark recognition/optical character recognition and crowd sourcing

Vital sign	COVID positive (188)	COVID negative (1960)
	Mean	
Systolic BP (mmHg)	126.95	128.46
Diastolic BP (mmHg)	78.90	79.42
Heart rate	88.57	86.47
Respiratory rate	16.55	16.84
Temperature (°F)	98.90[a]	98.48
O_2 saturation (%)	97.70	97.87
Symptom	**Number of patients with symptom (% of total)**	
Cough	165 (87.7%)	1,712 (87.4%)
Shortness of breath	70 (37.2%)	966 (49.2%)[a]
Reported fever		
Subjective	52 (27.7%)[a]	519 (26.5%)
100.4°F–102°F	58 (30.9%)[a]	290 (14.8%)
>102°F	9 (4.8%)[a]	44 (2.2%)
Suspected or positive COVID contact	73 (38.8%)[a]	503 (25.7%)

Note: Data collected from intake forms were separated into those that tested positive or negative for COVID-19 and values were summarized either as mean values (systolic BP, diastolic BP, heart rate, respiratory rate, temperature, and O_2 saturation) or raw counts with percentages (presence of cough, shortness of breath, reported fever with ranges, or COVID-19 contacts). Values between the COVID-19 positive and negative groups were analyzed for statistical significance by using Mann–Whitney U tests or Fisher's exact tests.
[a]Statistical significance ($p < 0.05$).

A total of 2,814 forms were scanned, uploaded, and analyzed, 483 of which were duplicated entries for interparticipant agreement analysis or those tested for SARS-CoV-2 multiple times. Additionally, 87 records were the incorrect format, 38 entries were incidentally repeated by the same participant, and 58 patients were excluded for age younger than 18 years. This yielded a total of 2,148 unique records, 188 (8.8%) of which were SARS-CoV-2 positive on PCR and 1,960 (91.2%) of which were SARS-CoV-2 not detected on PCR. Common symptoms and vital signs are displayed (►Table 2) with temperature (°F), shortness of breath (SOB), reported fever, and suspected or positive COVID contact, all showing statistical significance between detected and nondetected subpopulations. All had greater values in the SARS-CoV-2 detected subpopulation with the exception of proportion of patients reporting SOB, which was greater in the nondetected subpopulation.

Discussion

The primary goal of this project was to demonstrate a novel approach to efficiently capturing large amounts of clinical data through paper-based assessment forms to expedite triage, which was accomplished with a novel approach that combined (1) OMR/OCR tools to convert paper forms, (2) validating extracted data with crowdsourcing, and (3) a custom user interface for quick data review. However, there are other benefits to this system in the setting of a pandemic. When managing infectious disease risks, health care workers must be especially cautious about interacting with equipment that may carry fomites, such as clinical workstations which can carry SARS-CoV-2 for up to 3 days.[25,26] Complex workflows for health care workers wearing full PPE while assessing potentially-infected patients may increase this risk. Having access to a convenient and low-touch method for clinical documentation has value. Using preformatted paper forms for clinical documentation with subsequent OMR/OCR technology for data capture offers one solution that allows health care workers to assess patients efficiently without contaminating computer workstations or breaching their PPE.

This paper-based assessment has additional benefits beyond worker safety. First, the approach has the benefit of enforcing templated assessment; limiting the document to the most common clinical features of COVID-19 makes assessments more focused, efficient, and reproducible. For those patients who are not currently patients within our institution's health care system and lack a medical record, a paper form saves even more valuable time as the workflow is not interrupted while awaiting creation of a new medical record. Using paper, which is single-use and less viable for the virus, carries a much lesser infection risk.[25] Moreover, once scanned, the document can be destroyed.

Previous studies have demonstrated the accuracy of using paper triage forms for the purpose of screening in clinics and nonemergent situations.[5,10,11,27] Our addition of human verification as a backup validation step for the OMR/OCR served two functions: it allowed for quality control of a system that can be prone to physician error and allowed the capture of data that might not fit within the preset prompts.[28] Given the large volume of assessments during the initial COVID-19 spike, human validation needs were higher than normal.

This project applied crowdsourcing to improve validation efficiency. Crowdsourcing has been used in the past for a variety of purposes.[29] Primary issues regarding crowdsourcing annotation of sensitive patient information have been surrounding the qualification of the "crowd" and patient privacy.[30] The current COVID-19 pandemic has displaced a

large number of trained medical personnel and medical students from normal clinical participation. This greatly increased the pool of qualified annotators.

Additionally, in comparison to prior studies using double-data entry by two independent evaluators, this study achieved comparable agreement between reviewers in a comparable time (1.1 seconds per field vs. 1.1 seconds per field in this study).[31] In comparison to the reported correct recognition rates of existing OMR/OCR technologies, our OMR system was not quite as accurate (OMR 70%; 92.4 and 98.6%; and >99% correct recognition), but the addition of the crowdsourced validation sufficiently limited the potential for errors.[5,10,11,32] However, the studies above both applied OMR/OCR systems to fewer forms ($n = 221$, $n = 398$) and with fewer fields.[11,32] In addition, in the Biondich studies, up to 43% of values were analyzed by the OMR system required manual validation.[10] Given the timeline of COVID-19 and need to gather data quickly, the combination of a rapidly developed novel OMR/OCR tool with crowdsourced validation was a unique and efficient way to validate a large amount of data quickly.

The OMR/OCR tool is able to accurately fill in a large proportion of fields, thus reducing the manual effort needed from the crowd workers. While not 100%, we believe this extraction starting point allows workers to shift from a data extract task to a data validation task. Our observation from reviewing many intake forms is that the validation process is much faster than the data extract process. Intuitively, it is much easier to look across the scanned form, see which fields are marked, and verify that they are also captured. Moreover, because of the mirrored UI, it is readily apparent which forms have been accurately extracted by the tool versus those that need correction.

This study's findings should be considered in light of its limitations. First, it was necessary to include narrative-text fields within the paper form to allow providers to accurately and completely document potential COVID-19 cases. Because of this, basic OMR could not be used to recognize all aspects of the paper form. Future versions of the system may incorporate methods for numeric text identification and later handwritten text. Second, despite the OMR automation of checkbox components, crowdsourced human verification was required. This necessitated analysis of participant agreement to ensure validity. While this research team did consider alternative open source and commercial automated paper form processing systems, the setup time required as well as the potential legal obstacles involving protected health information (PHI) led the team to design and deploy their own system.

The four common categories of error that can occur when health care workers utilize paper forms include errors of omission or commission by the physician or system. Examples include leaving components or categories blank (error of omission by physician) or if the tools necessary to complete a category were not available (error of omission due to systems).[4] Physician writing in the margin outside of components (error of commission by the physician) is also troublesome when using paper forms. Potential strategies to decrease errors of commission such as this could include brief training videos. Another limitation was the balance between having boxes that were of adequate size for physician input and OMR recognition while containing all information on a single sheet to maintain portability and efficiency.

Further study is needed to fully evaluate the various technological components presented in this work. First, future work will analyze the accuracy of the OMR algorithms and the algorithms' agreement with crowd participants. Second, future work will examine the extent to which the presented UI impacts data validation times compared with simpler user interfaces. While these additional studies are necessary, this work outlines the end-to-end design, capture, extraction, and storage of COVID-19 assessment information.

Conclusion

This work describes the use of a paper triage form used to collect data during in-person evaluations of COVID-19, which was analyzed by using a novel optical mark/character recognition technology and validated by crowdsourced health care workers. Using a novel user interface, human validators were able to quickly correct errors and validate the results of the OMR. These data demonstrate that the combination of OMR technology and crowdsourced human validation can expedite data transformations from paper documents to electronic information during a pandemic.

Clinical Relevance Statement

When rapidly triaging patients as was done during the heights of the COVID-19 pandemic, it is significantly easier for health care providers to use paper triage forms than electronic records. However, one drawback of the paper forms is that patient information will have to be transferred to their electronic record as some point, and this delays the process. Here, we have demonstrated that using both OMR/OCR technology as well as crowdsourcing can expedite the transfer of patient information from paper forms to electronic records in the setting of a pandemic.

Multiple Choice Questions

1. SARS-CoV-2 has been shown to be viable on clinical workstations for up to:
   a. 1 day
   b. 3 days
   c. 5 days
   d. 7 days

   Correct Answer: The correct answer is option b. SARS-CoV-2 is the positive-sense single-stranded RNA virus that is responsible for COVID-19. It has shown to have transmissibility via fomites, as a study published in New England Journal of Medicine identified its viability on several surfaces including cardboard, stainless steel, and plastic, and found that the virus had a maximum viability of 3 days on plastic surfaces.[23] As clinical workstations are largely made of plastic, this is one advantage that paper triage forms have over directly inputting into EMR during

a viral pandemic. 1 day, 5 days, and 7 days are the wrong answers as they do not accurately show the viability of SARS-CoV-2 on plastic surfaces.

2. When physicians leave a category or section blank on paper triage forms, this is best called an error of:
   a. Commission
   b. Poor patient care
   c. Omission
   d. Judgment

   Correct Answer: The correct answer is option c. Medical errors are often classified into two categories: errors of omission and errors of commission. Errors of omission occur as a result of action not taken, exemplified in this manuscript by leaving a section or part blank on a triage form. This is the correct answer. Errors of commission occur as a result of the wrong action taken, exemplified here as writing outside the margins on a paper triage form such that OMR/OCR technology cannot identify the information. Errors of poor patient care and judgment do not fit this scenario as well as omission.

3. OMR, in the context of this paper, stands for:
   a. Optional medical resuscitation
   b. Omission of medical records
   c. Oversharing of medical records
   d. Optical mark recognition

   Correct Answer: The correct answer is D. Optical mark recognition describes the process of capturing human-marked data from paper forms. It is a crucial aspect behind the transfer of data from paper documents to electronic records and was utilized in this manuscript to rapidly transfer COVID data. The other three options, optional medical resuscitation, omission of medical records, and oversharing of medical records, do not accurately describe the most commonly used definition of OMR.

Protection of Human and Animal Subjects

None.

Funding

None.

Conflict of Interest

None declared.

References

1 Patel PD, Cobb J, Wright D, et al. Rapid development of telehealth capabilities within pediatric patient portal infrastructure for COVID-19 care: barriers, solutions, results. J Am Med Inform Assoc 2020;27(07):1116–1120
2 Kim SI, Lee JY. Walk-through screening center for COVID-19: an accessible and efficient screening system in a pandemic situation. J Korean Med Sci 2020;35(15):e154
3 Islam MS, Rahman KM, Sun Y, et al. Current knowledge of COVID-19 and infection prevention and control strategies in healthcare settings: a global analysis. Infect Control Hosp Epidemiol 2020;41(10):1196–1206
4 Ferioli M, Cisternino C, Leo V, Pisani L, Palange P, Nava S. Protecting healthcare workers from SARS-CoV-2 infection: practical indications. Eur Respir Rev 2020;29(155):200068
5 Downs SM, Carroll AE, Anand V, Biondich PG. Human and system errors, using adaptive turnaround documents to capture data in a busy practice. AMIA Annu Symp Proc 2005;2005:211–215
6 Collen MF. Clinical research databases–a historical review. J Med Syst 1990;14(06):323–344
7 Shah NH, Tenenbaum JD. The coming age of data-driven medicine: translational bioinformatics' next frontier. J Am Med Inform Assoc 2012;19(e1):e2–e4
8 Bhargava BK, McDonald CJ, Rivera HP, McCarthy LJ, Blevins L. Development and Implementation of a Computerized Clinical Laboratory System. Lab Med 1976;7(12):28–37
9 Tafti AP, Baghaie A, Assefi M, Arabnia HR, Yu Z, Peissig P. OCR as a Service: An Experimental Evaluation of Google Docs OCR, Tesseract, ABBYY FineReader, and Transym. In: Bebis G, Boyle R, Parvin B, et al., eds. Advances in Visual Computing. Lecture Notes in Computer Science. Springer International Publishing; 2016: 735–746
10 Biondich PG, Overhage JM, Dexter PR, Downs SM, Lemmon L, McDonald CJ. A modern optical character recognition system in a real world clinical setting: some accuracy and feasibility observations. Proc AMIA Symp 2002:56–60
11 Biondich PG, Anand V, Downs SM, McDonald CJ. Using adaptive turnaround documents to electronically acquire structured data in clinical settings. AMIA Annu Symp Proc 2003; 2003:86–90
12 Shiffman RN, Brandt CA, Freeman BG. Transition to a computer-based record using scannable, structured encounter forms. Arch Pediatr Adolesc Med 1997;151(12):1247–1253
13 Titlestad G. Use of document image processing in cancer registration: how and why? Medinfo 1995;8(Pt 1):462
14 Bussmann H, Wester CW, Ndwapi N, et al. Hybrid data capture approach for monitoring patients on highly active antiretroviral therapy (HAART) in urban Botswana. Bull World Health Organ Int J Public Health 2006;842:127–131
15 Bergeron BP. Optical mark recognition. Tallying information from filled-in 'bubbles'. Postgrad Med 1998;104(02):23–25
16 Shiffman R, Brandt C, Hoffman M, Wiig W, Fernandes L. SEURAT: scanned entry of structured data for a pediatric health maintenance record system. Accessed April 18, 2020 at: https://www.researchgate.net/publication/25901454_SEURAT_Scanned_Entry_of_Structured_Data_for_a_Pediatric_Health_Maintenance_Record_System
17 Loke SC, Kasmiran KA, Haron SA. A new method of mark detection for software-based optical mark recognition. PLoS One 2018;13(11):e0206420
18 Chouvatut V, Prathan S. The flexible and adaptive X-mark detection for the simple answer sheets. 2014 International Computer Science and Engineering Conference. Accessed 2014 at: https://ieeexplore.ieee.org/document/6978236
19 Sattayakawee N. Test scoring for non-optical grid answer sheet based on projection profile method. Int J Inf Educ Technol 2013: 273–277
20 Rakesh S, Atal K, Arora A. Cost effective optical mark reader. Int J Comput Sci Artif Intell Accessed April 18, 2020 at: https://scholar.google.com/scholar_lookup?journal-=International+Journal+of+Computer+Science+and+Artificial+Intelligence&title=Cost+effective+optical+mark+reader&author=S+Rakesh&author=K+Atal&author=A+Arora&volume=3&publication_year=2013&pages=44&
21 Bradski G. The Open CV Library. Dr Dobbs J Softw Tools Accessed 2000 at: https://www.drdobbs.com/open-source/the-opencv-library/184404319
22 Ye C, Coco J, Epishova A, et al. A crowdsourcing framework for medical data sets. AMIA Jt Summits Transl Sci Proc 2018; 2017:273–280

23 Harris PA, Taylor R, Thielke R, Payne J, Gonzalez N, Conde JG. Research electronic data capture (REDCap)–a metadata-driven methodology and workflow process for providing translational research informatics support. J Biomed Inform 2009;42(02): 377–381
24 Harris PA, Taylor R, Minor BL, et al;REDCap Consortium. The REDCap consortium: building an international community of software platform partners. J Biomed Inform 2019;95:103208
25 van Doremalen N, Bushmaker T, Morris DH, et al. Aerosol and surface stability of SARS-CoV-2 as compared with SARS-CoV-1. N Engl J Med 2020;382(16):1564–1567
26 Popescu S. Roadblocks to infection prevention efforts in health care: SARS-CoV-2/COVID-19 response. Disaster Med Public Health Prep 2020;14(04):538–540
27 Anand V, Carroll AE, Downs SM. Automated primary care screening in pediatric waiting rooms. Pediatrics 2012;129(05):e1275–e1281
28 Fifolt M, Blackburn J, Rhodes DJ, et al. Man versus machine: comparing double data entry and optical mark recognition for processing CAHPS survey data. Qual Manag Health Care 2017;26 (03):131–135
29 Leung GM, Leung K. Crowdsourcing data to mitigate epidemics. Lancet Digit Health 2020;2(04):e156–e157
30 Spasic I, Nenadic G. Clinical text data in machine learning: systematic review. JMIR Med Inform 2020;8(03):e17984
31 Kawado M, Hinotsu S, Matsuyama Y, Yamaguchi T, Hashimoto S, Ohashi Y. A comparison of error detection rates between the reading aloud method and the double data entry method. Control Clin Trials 2003;24(05):560–569
32 Paulsen A, Overgaard S, Lauritsen JM. Quality of data entry using single entry, double entry and automated forms processing–an example based on a study of patient-reported outcomes. PLoS One 2012;7(04):e35087

