Published in final edited form as:

Acad Radiol. 2013 February ; 20(2): 238–242. doi:10.1016/j.acra.2012.09.016.

Unbiased Review of Digital Diagnostic Images in Practice: Informatics Prototype and Pilot Study

Anthony F. Fatenos, MD, PhD*, Nabile M. Safdar, MD*, Paul G. Nagy, MD, Reuben Mezrich, MD, PhD, and Jonathan S. Lewin, MD

Russell H. Morgan Department of Radiology and Radiological Science, Johns Hopkins University, 600 North Wolfe Street, Baltimore, Maryland 21287 (A.F.F., P.G.N., J.S.L.); Sheikh Zayed Institute of Pediatric Surgical Innovation, Department of Diagnostic Imaging, Children's National Medical Center, Washington DC (N.M.S.); Department of Diagnostic Radiology and Nuclear Medicine, University of Maryland School of Medicine, Baltimore, MD (R.M.)

Abstract

Rationale and Objectives—Clinical and contextual information associated with images may influence how radiologists draw diagnostic inferences, highlighting the need to control multiple sources of bias in the methodological design of investigations involving radiological interpretation. In the past, manual control methods to mask review films presented in practice have been used to reduce potential interpretive bias associated with differences between viewing images for patient care versus reviewing images for purposes of research, education, and quality improvement. These manual precedents from the film era raise the question whether similar methods to reduce bias can be implemented in the modern digital environment.

Materials and Methods—We built prototype “CreateAPatient” information technology for masking review case presentations within our institution’s production Radiology Information and Picture Archiving and Reporting Systems (RIS and PACS). To test whether CreateAPatient could be used to mask review images presented in practice, six board-certified radiologists participated in a pilot study. During pilot testing, seven digital chest radiographs, known to contain lung nodules and associated with fictitious patient identifiers, were mixed into the routine workload of the participating radiologists while they covered general evening call shifts. We tested whether it was possible to mask the presentation of these review cases, both by probing the interpreting radiologists to report detection and by conducting a forced-choice experiment on a separate cohort of 20 radiologists and information technology professionals.

Results—None of the participating radiologists reported awareness of review activity, and forced-choice detection was less than predicted at chance, suggesting radiologists were effectively blinded. In addition, we identified no evidence of review reports unsafely propagating beyond their intended scope or otherwise interfering with patient care, despite integration of these records within production electronic workflow systems.

Conclusion—Information technology can facilitate the design of unbiased methods involving professional review of digital diagnostic images.

Keywords

true screening; in-service monitoring; blind review; objective determination; unannounced standardized patients

Address correspondence to: A.F.F. Anthony.Fatenos@jhmi.edu, phone (410) 929-1247, fax (516) 224-3660.

*Co-first authors

INTRODUCTION

Clinical and contextual information associated with images may influence how radiologists draw diagnostic inferences. For example, in mammography, when a regularly screened woman receives a diagnosis of breast cancer without suspicion having been raised on her last screening mammogram, a quality-related question naturally arises: Was the last screening mammogram truly negative? Multiple studies of these positive “interval cancer” mammograms suggest that the answer to this question is complex. As an extreme illustrative example, one reviewer in a recent study reported positive findings in 3 of 20 interval-cancer mammograms when they were de-identified and presented mixed with normal mammograms, yet reported positive findings in 17 of the same 20 mammograms when they were unmixed and associated with additional clinical information. When aggregating this data statistically, the magnitude of design effects remains significant and large, with positive finding rates ranging over an order of magnitude across review conditions in most studies (1).

The importance of review design is not limited to studies of interval breast cancer. Semelka and colleagues (2) recently reported that out of 31 external radiologist reviewers, none documented suspicion of spinal instability when reviewing a CT showing 1-mm symmetric facet widening at T10. The study CT was presented with five control CTs in the context of a quality assurance review. Study reviewers were not informed that the CT had also been the subject of a settled lawsuit, in which four out of four expert witnesses for the plaintiff had represented the spinal instability findings as expected per standard of care.

In general, the literature on image review methodology suggests that radiologists incorporate all available information associated with an image, including explicit and implicit contextual information, in order to reduce the uncertainty of diagnostic reporting. Specifically, the degree to which radiologists suspect findings informs the likelihood of reporting them, and review designs influence perceived probabilities (3,4). How can this confounding effect be handled in experiments involving radiological interpretation? Two basic strategies are available: 1) accept bias as a given, and aim to control it by comparable allocation among reviewers (controlled reviews, with an emphasis on internal validity, for example 5); or 2) aim to minimize bias by studying image review in practice (observational reviews, with an emphasis on external validity, for example 6).

Interestingly, these design strategies need not be mutually exclusive. We are aware of two examples of controlled image review in which radiologists were observed in practice (7,8). In both studies, the film and paper records of selected patients were manually controlled and included blinding radiologists to the presentation of review cases. The logistical requirements for such unbiased designs were considerable, requiring manipulation of paper charts and films in order to mask potential sources of bias, such as awareness of image review, other reviewers’ reports, and patients’ clinical outcomes. Reflecting on these challenges, Gordon and colleagues (8) wrote in their conclusion, “We cannot comment on the logistics of performing a similar study in a digital environment...However, where it is feasible, we believe it will improve the validity of review results and provide a context more similar to the one in which the original radiologist had to work” (see also 9,10).

This question, whether unbiased image review is feasible in a digital environment, motivates the present study. In particular, we set out to design, develop, and deploy a prototype for presenting masked images within production Radiology Information and Picture Archiving and Communication Systems (RIS and PACS, review masking pilot). We also tested the specific hypothesis that our prototype would lead to observer blinding with a failure rate no

greater than 5%, a standard target for related studies of unannounced standardized patients (forced-choice detection test, 11).

MATERIALS AND METHODS

Approval to present faculty radiologists with masked review cases was obtained, subject to the following restrictions: 1) Participating radiologists provided consent as a condition for research eligibility, understanding that the presentation of review cases could be masked anytime over a 12-month period of practice; 2) No more than two review cases were to be presented per radiologist per 8-hour shift; 3) Each review presentation was monitored in real time via computer by the study administrator in order to guarantee the integrity of the hospital's information systems and prevent non-consensual study participation.

Review Masking Pilot

Six board-certified, faculty radiologists, two each from the musculoskeletal, emergency, and body divisions, agreed to participate and were presented with masked review chest radiographs on evening shifts ranging between seven and nine hours in duration. Seven chest radiographs were selected from a set obtained from adult patients participating in a previously detailed study of lung nodule detection, with CT of the chest used as a reference standard (12). The seven radiographs contained visible lung nodules ranging from 6 mm to 18 mm in greatest axial dimension on CT. The digital radiographs were identified exclusively by association with electronic metadata and contained no burned-in patient identifiers.

Figures 1 and 2 demonstrate the informatics architecture and workflow we designed to integrate the presentation of masked review cases within the hospital's RIS reporting application (GE, Fairfield, Connecticut) and PACS (Agfa, Mortsel, Belgium). Electronic medical record (EMR) integration was deferred for this study, since masked review cases were presented on a teleradiology worklist, with each patient's clinical notes and lab values residing outside the radiologist's locally accessible EMR.

Figures 1 and 2 illustrate a server-based, image-reproduction model for presenting masked review chest radiographs. Specifically, we created a database to store research information about review cases distributed within the PACS and RIS reporting application. Review cases were reproductions of original cases, consisting of exact pixel copies of patients' original radiographs, but associated with the following manufactured set of textual metadata: first name, middle initial, last name, date of birth, account number, sex, address, phone number, attending physician, referring physician, study accession number, medical record number, and department number. None of the manufactured identifying information matched the patient, except for sex, zip code, and year of birth, since radiologists might use these group-level identifiers to inform their reports.

A research application was built to serve as the front-end of the research database and to facilitate masked case presentation. We named the application "CreateAPatient." It consisted of a web browser-based user interface for entering manufactured metadata, as well as logic for translating user input into a machine-readable format and for routing the data to the PACS and RIS reporting application. CreateAPatient incorporated the freely available software DicomWorks (Villeurbanne, France) and dcm4chee (Milwaukee, Wisconsin) to enable adherence to the Digital Imaging and Communications in Medicine (DICOM) standard for communication with the PACS. The freely available application Mirth Connect (Irvine, California) was used to adhere to the Health Level 7 (HL7) standard for communication with the RIS reporting application.

We prevented potentially corrupting communication between the RIS reporting application, the RIS database, and other connected electronic systems by capitalizing on error-checking functionality in the communication protocol of our RIS. Specifically, we manufactured accession numbers over a shorter digit range, such that the RIS database rejected them. We prevented potentially disruptive radiologist-to-referrer communication by use of a virtual answering service, with callbacks provided by a confederate of the study administrator.

Forced-choice detection test

Radiologists participating in the blinded review pilot were instructed to notify the study administrator if they suspected any of their cases were reviews. As an additional test of masking effectiveness, we recruited a separate cohort of 10 radiologists and 10 radiology IT professionals to participate in a forced-choice detection test. Test participants were presented with a live worklist containing approximately 20 requests for the interpretation of chest radiographs. Participants were instructed that one of the listed studies referred to a masked review case. Participants were requested to make three attempts to identify the review case.

This study was approved by our medical center's Institutional Review Board (IRB).

RESULTS

Blinded review pilot

As a prototype for presenting masked review images in routine radiological practice, we built working software based on the architecture and workflow shown in Figures 1 and 2. This CreateAPatient software included a database and user interface, with customized communication interfaces to the hospital's PACS and RIS reporting application. Figure 3 demonstrates the user interface of the CreateAPatient application.

Real-time monitoring and follow-up demonstrated no evidence of review records propagating beyond the intended research scope. In particular, no record of a procedure or result associated with a review case corrupted the EMR, confused a non-participating radiologist or clinician, or interfered with the work of downstream billing or oversight personnel at the referring or interpreting hospitals.

Radiologist-case interactions involved six radiologists and seven review radiographs. The mean radiologist-case interaction time, measured from case presentation to report signing, was 77 minutes (range 29–161). Three of the seven radiographs containing lung nodules were identified as containing lung nodules in the diagnostic review reports (sensitivity 43%, exact 95% confidence interval 10–82%). Two of the three radiologists who identified these nodules also called the manufactured phone number of the referring clinician.

Forced-choice detection test

None of the six radiologists participating in the presentation pilot study reported detecting a masked review case. Of the 20 participants in our separate forced-choice detection test, one (5%) accurately identified the manufactured order for a review case when it was embedded on a live worklist of approximately 20 original orders.

DISCUSSION

In the present study, we have demonstrated the feasibility of an informatics design for unbiased image review. Studies of film review have shown that unblinded designs in artificial settings can yield dramatically skewed estimates of diagnostic accuracy, motivating

our present effort to design a potentially scalable prototype for unbiased image review in the digital era.

Scope and Terminology

This study is informed by the literature on film review for interval and incident breast cancer (7,8), radiological expert witnesses (2), blinded pathology (13), unannounced standardized patients in practice (11), A/B computer-user experiments (14), and covert testing and threat image projection at airports (15). This multidisciplinary corpus is related by a common empirical design: the review of investigator-controlled cases by observers in practice with study information withheld to minimize bias. In the diagnostic imaging literature, this design theme has been described using keywords such as “blind review” (7), “true screening” (8), “in-service monitoring” (9), “objective determination” (2), and “true blind review” (10). In the standardized patient literature, keywords have included “pseudopatients” (16), “incognito standardised patients” (17), “stealth standardised patients” (18), and “unannounced standardised patients in real practice” (11). This variability in the medical terminology for unbiased review limits multidisciplinary exchange and increases the risk that we may remain unaware of some relevant prior literature.

In addition to variable terminology, the use of the keyword “blinding” for bias reduction may also cause confusion, since investigators use the same keyword for withholding variable quantities of identification from reviewers. For example, at one extreme, Gordon and colleagues describe an accidental design in which mammography technologists privately masked the paper and film records of interval cancer patients, so that radiologist reviewers only became aware of their own identity as research subjects after the fact. The present review design, as well as that of de Rijke and colleagues (7), belongs one notch down on this blinding hierarchy, since radiologist reviewers knew of their own subject identity, but were not informed when they switched from the role of radiologist to the role of research reviewer. Alternatively, we might have used our CreateAPatient application to mask patient identity not by using the names of fictitious individuals, but rather by using a generic mask (eg, “Review Patient”), in which case radiologists would have known when they switched to a reviewing role, but would have remained blinded to patient and diagnostic identity. Moving further down this hierarchy, multiple review studies of interval cancer in mammography have employed “blinded, mixed” reviews, in which radiologists play the reviewer role in an artificial setting continuously, but are nevertheless blinded to patient and diagnostic identity by the presentation of de-identified positive and negative cases (1,19). Finally, at the opposite extreme, fully unblinded reviews occur when images are interpreted under conditions of certainty regarding patient diagnosis and outcome.

To address a final potential point of confusion, it might be worth noting that “blinding” in its most familiar medical use to describe therapeutic trial (not review) design (ie, “double-blind randomized controlled trial”) has evolved specifically to refer to withholding the identity of interventions from research subjects and/or investigators when two or a few interventions are randomly allocated one-to-one among many subjects (20). To avoid confusion, we advise remembering that a “blinded study” means only that study information has been deliberately withheld to reduce bias, but exactly what information is withheld and why should be separately specified.

Study Limitations and Future Direction

The main contribution of this study is the specification of an informatics architecture for implementing unbiased digital image review within a production PACS and RIS system. Principal limitations are the small sample size and associated narrowly focused feasibility aim. In the future, we look forward to applying the methods developed here for statistically

powered trials of diagnostic outcomes. As part of a methodologically focused and foundational proof of concept, however, the feasibility of masking digital review cases is presently supported by the absence of detection reported among the six radiologist reviewers in our pilot. In addition, fewer than 5% of expert observers in the forced-choice test (fewer than expected by chance) were able to distinguish masked from new cases, when these were mixed on the same worklist. We are not aware of blinded image reviews that have quantified detection rates. A wide range, between 0% and 70%, has been reported for the detection of USPs by clinicians in practice (11).

The CreateAPatient software was limited by lack of EMR integration. Consequently, our pilot was restricted to worklists on which radiologists have no expectation of looking patients up in the EMR. For many review applications, however, it may be sufficient to blind reviewers to case identity (patient = “Review Patient” not “Joe Smith”), in which case EMR-masking technology would not be essential. Another limitation of CreateAPatient is that manufacturing accession numbers over a shorter digit range to prevent downstream record propagation may not generalize to all RIS manufacturers or robustly serve to de-identify review images in the future.

In conclusion, our results encourage further development of software for experimental control over the presentation of diagnostic imaging information in practice, especially given the potential of such software to scale in response to demands for unbiased evidence in this era of increasingly digital and evidence-based medicine.

Acknowledgments

We wish to thank Sean Berenholtz, John Bridges, Randy Buckner, Philip Cascade, Mythreyi Chatfield, Andrew Crabb, Daniel Durand, Richard Gunderman, Marc Edwards, Edi Karni, David Larson, Harold Lehmann, Jesse Mazer, Rodney Owen, Mike Kraut, Lyle Malotky, Eric Maskin, Daniel Marcus, David Newman-Toker, Peter Pronovost, Eliot Siegel, Ken Wang, Siren Wei, David Yousem, and Samuel Yousem for stimulating discussions of review methodology.

Grant Support: GE-Radiology Research Academic Fellowship (GERRAF).

References

1. Ciatto S, Catarzi S, Lamberini MP, et al. Interval breast cancers in screening: the effect of mammography review method on classification. *Breast*. 2007; 16(6):646–52. [PubMed: 17624779]
2. Semelka RC, Ryan AF, Yonkers S, et al. Objective determination of standard of care: use of blind readings by external radiologists. *AJR Am J Roentgenol*. 2010; 195(2):429–31. [PubMed: 20651200]
3. Wolfe JM, Horowitz TS, Van Wert MJ, et al. Low target prevalence is a stubborn source of errors in visual search tasks. *J Exp Psychol Gen*. 2007; 136(4):623–38. [PubMed: 17999575]
4. Gunderman RB. Biases in radiologic reasoning. *AJR Am J Roentgenol*. 2009; 192(3):561–4. [PubMed: 19234247]
5. Ryan A, Semelka R, Molina P. Evaluation of radiologist interpretive performance using blinded reads by multiple external readers. *Invest Radiol*. 2010; 45(4):211–6. [PubMed: 20177390]
6. Liu PT, Johnson CD, Miranda R, et al. A reference standard-based quality assurance program for radiology. *J Am Coll Radiol*. 2010; 7(1):61–6. [PubMed: 20129274]
7. de Rijke JM, Schouten LJ, Schreutelkamp JL, et al. A blind review and an informed review of interval breast cancer cases in the Limburg screening programme, the Netherlands. *J Med Screen*. 2000; 7(1):19–23. [PubMed: 10807142]
8. Gordon PB, Borugian MJ, Warren Burhenne LJ. A true screening environment for review of interval breast cancers: pilot study to reduce bias. *Radiology*. 2007 Nov; 245(2):411–5. [PubMed: 17848684]

9. Evertsz, C. Methods and system for in-service monitoring and training for a radiologic workstation. US Patent #7366676. 2008.
10. Glyptis P, Givens D. Developing a true blind review. ACR bulletin. 2011 Sep.189(3):24.
11. Rethans J-J, Gorter S, Bokken L, et al. Unannounced standardised patients in real practice: a systematic literature review. Med Educ. 2007; 41(6):537–49. [PubMed: 17518833]
12. Siegel E. Effect of dual-energy subtraction on performance of a commercial computer-assisted diagnosis system in detection of pulmonary nodules. Proc SPIE. 2005; 5748:392–8.
13. Renshaw, Aa; Pinnar, NE.; Jiroutek, MR.; Young, ML. Blinded review as a method for quality improvement in surgical pathology. Arch Pathol Lab Med. 2002; 126(8):961–3. [PubMed: 12171496]
14. Tang, D.; Agarwal, A.; Brien, DO., et al. Overlapping experiment infrastructure: more, better, faster experimentation categories and subject descriptors. Proceedings of the 16th ACM SIGKDD international conference on knowledge discovery and data mining; Washington, DC: ACM; 2010. p. 17-26.
15. Wetter, O.; Hardmeier, D. Covert testing at airports: exploring methodology and results. Security Technology; 2008.
16. Rosenhan D. On being sane in insane places. Science. 1973; 179:250–8. [PubMed: 4683124]
17. Maiburg BHJ, Rethans J-JE, van Erk IM, et al. Fielding incognito standardised patients as “known” patients in a controlled trial in general practice. Med Educ. 2004; 38(12):1229–35. [PubMed: 15566533]
18. Ozuah PO, Reznik M. Residents’ asthma communication skills in announced versus unannounced standardized patient exercises. Ambul Pediatr. 2007; 7(6):445–8. [PubMed: 17996838]
19. Hofvind S, Skaane P, Vitak B, et al. Influence of review design on percentages of missed interval breast cancers: retrospective study of interval cancers in a population-based screening program. Radiology. 2005; 237(2):437–43. [PubMed: 16244251]
20. Kaptchuk TJ. Intentional ignorance: a history of blind assessment and placebo controls in medicine. Bull Hist Med. 1998; 72(3):389–433. [PubMed: 9780448]

Figure 1.

Review masking system components and their connections. Arrows indicate directionality of information flow. PACS = Picture Archiving and Communication System, RIS = Radiology Information System, DB = database.

Figure 2.

Review masking workflow. Swimlane diagram illustrating process steps for radiologist reviewer and study administrator, with time flowing from top to bottom. Interpretations for masked review cases were blocked from entering the RIS DB. PACS = Picture Archiving and Communication System, RIS = Radiology Information System. DB = database.

The user interface for the CreateAPatient research application consists of three main components:

- Patient Information Form (Left):** A grey rectangular form containing various input fields for patient data. The fields are organized as follows:
 - Top row: First Name (text input), Middle Initial (text input).
 - Second row: Last Name (text input), Date of Birth (dropdown menu).
 - Third row: Account Number (text input), Sex (dropdown menu).
 - Fourth row: Address (text input), Phone Number (text input).
 - Fifth row: Attending (text input).
 - Sixth row: Referring Physician (text input).
 - Seventh row: Accession Number (text input).
 - Eighth row: MRN (text input).
 - Ninth row: Department Number (text input).At the bottom of the form are three buttons: "Submit to HL7", "Submit to Dicom", and "Reset".
- HL7 Viewer (Top Right):** A light blue rectangular window with a title bar "HL7 Viewer". It contains a large white area for displaying HL7 data, with a vertical scrollbar on the right side.
- Dicom Viewer (Bottom Right):** A light blue rectangular window with a title bar "Dicom Viewer". It contains a large white area for displaying Dicom data, with a vertical scrollbar on the right side.

Figure 3.
User interface for the CreateAPatient research application.