



White Paper



# Diagnostic Accuracy in Radiology: Defining a Literature-Based Benchmark

## Contents

<b>Highlights</b>	<b>3</b>
<b>Introduction</b>	<b>3</b>
<b>Materials &amp; Methods</b>	<b>4</b>
<b>Results</b>	<b>5</b>
<b>Discussion</b>	<b>6</b>
<b>Conclusion</b>	<b>8</b>

## Highlights

**Purpose:** All Physicians make diagnostic errors in treating their patients, and even well-trained and diligent radiologists are not exempt from making errors. Determining the expected prevalence of errors, or a so-called “acceptable” error rate, is a difficult task because there are strong disincentives for error reporting. Outside of mammography, there are no agreed-upon industry standards establishing desirable goals for diagnostic accuracy or consistent frameworks for how to measure errors in interpretation. The intent of this paper is to establish a benchmark standard for interpretive accuracy rates based on a review of the available literature concerning interpretive errors.

**Materials & Methods:** There are hundreds of research articles reviewing error rates in radiology. For this review, we narrowed our tabulation of results to a manageable number of six research papers, which collectively reported on the error rate of approximately 650,000 examinations. From these research papers, we believe that a reasonable benchmark for accuracy may be extrapolated for the practice of radiology. The additional hundreds of papers present similarly-reported error rates.

**Results:** Based upon double-blind interpretations of studies from different modalities, interpreted at community hospitals and universities, the blended error rate for a wide range of modalities is 4.4%, with a possible range of errors between 0.8% and 9.2% depending on the type of studies interpreted, modality mix and subspecialty expertise of the radiologist. General radiologists limiting their interpretations to x-rays, mammograms and ultrasounds will have a lower error rate of 3.48%. Computed tomography (CT) abdominal and pelvis interpretations, widely regarded as very complex and difficult, will have the highest disagreement rate between radiologists and clinicians, approaching 32%. The error rate for CT abdominal and pelvic examinations seen from patients through the emergency department will be as high as 7%.

**Conclusion:** While it is acknowledged that the lack of a consistent industry framework for error evaluation confounds this task, it is possible to draw meaningful conclusions about diagnostic errors in radiology.

*Based upon double-blind interpretations of studies from different modalities, interpreted at community hospitals and universities, the blended error rate for a wide range of modalities is 4.4%, with a possible range of errors between 0.8% and 9.2% depending on the type of studies interpreted, modality mix and subspecialty expertise of the radiologist.*

## Introduction

Autopsy series performed over decades have revealed diagnostic error rates between 4.1% and 49%. Perceptual specialties (e.g., radiology, pathology, and dermatology) have generally the lowest error rates, at less than 5%, while clinical specialties have diagnostic error rates ranging from 10% to 15%.<sup>1</sup> Diagnostic error rates from the emergency department (ED) ranged from 0.6% to 12%.<sup>2</sup> The Pennsylvania Patient Safety Authority reported, from the period from January 2005 through August 2009, that the top five categories for misdiagnoses by clinicians in all specialties were metastatic cancer (12%), fractures (4%), pulmonary embolism (4%), acute coronary syndrome (2%), and appendicitis (2%).

Most radiologists and groups do not analyze their errors, but rather correct errors as they are detected by referring physicians or other members of their own group. Radiologists often fail to tabulate errors, study this data, and develop methodologies for improvement. There is a lack of meaningful agreement on the real prevalence of diagnostic errors in radiology. Referring physicians with close friendships with radiologists do not long remember or hold their friends accountable for routine errors, resulting in a perception about error frequency and severity that is formed by the strength of the personal relationship, rather than the actual performance. It is typical to hear members of the hospital medical staff say, "Our radiologists never make errors." Unfortunately, this statement is not supported by any published research from major universities, community hospitals, large radiology groups, teleradiology organizations, or the American College of Radiology. FitsGerald<sup>3</sup> noted: "The lack of radiologic focus on error analysis may reflect our traditional medical culture which places a heavy emphasis on personal responsibility and autonomy of action. Mistakes should not be made, and if they are, they are indicative of personal and professional failure. Medicine lags behind safety cultures in other walks of life, e.g. aviation, in applying a systems approach to error. Such an approach is less concerned with who made the mistake but rather why the mistake was made and how it happened."

In 1959, L. Henry Garland<sup>4</sup> published the pioneering article on diagnostic errors in radiology, noting that a radiologist "missed" approximately 30% of positive findings. When other radiologists were confronted with this information, their response was, "Well, in my everyday work this does not apply; I would do better than those busy investigators." More than 60 years after the publication of his article, hundreds of research papers have been published that confirm an error rate in all areas of radiology, and indeed, in all areas of medicine. As noted by Seigle<sup>5</sup>, "Radiologists make a certain inescapable minimum number of interpretative errors, no matter how diligent they may be."

With the exception of mammography, no radiology professional organization has published standards on the acceptable rate of error or desirable goals for the accuracy of radiology interpretation.

This literature review of more than 100 articles pertaining to errors in radiology is an attempt to develop a "literature-based" benchmark for accuracy in radiology.

*"The lack of radiologic focus on error analysis may reflect our traditional medical culture which places a heavy emphasis on personal responsibility and autonomy of action."*

## Materials & Methods

There are hundreds of research articles reviewing error rates in radiology. For this review, we narrowed our tabulation of results to a manageable number of six research papers<sup>6,7,8,9,10</sup>, which collectively reported on the error rate of approximately 650,000 examinations. From these research papers, we believe that a reasonable benchmark for accuracy may be extrapolated for the practice of radiology. The additional hundreds of papers present similarly-reported error rates.

## Results

Numerous investigators have substantiated Garland's findings relative to x-ray findings of errors.<sup>11 12 13 14 15 16</sup> Herman et al. published that Harvard University radiologists disagreed on the interpretation of chest x-rays 56% of the time<sup>17</sup>, and recent studies have confirmed a 35% error rate among radiologists interpreting radiologic studies in trauma patients.<sup>18 19</sup>

It is important to define the term "error rate," so as to understand the difference between retrospective analysis of cases with known errors and error rates encountered in the daily interpretation of studies. Error rates can be calculated in two different ways, depending on what is used as the denominator. The following example of this is explained by Berlin.<sup>20</sup>

"If a series of 100 roentgenograms (x-rays) contains 10 positive and 90 negative films, and a reviewer misses three of the positive films and over-reads two of the negative films, he may be regarded as having only a 5% error. On the other hand, since the series of 100 roentgenograms is being examined to detect patients with disease, the reviewer who misses three of the 10 positive films has an error rate of 30%. Coupled with an over-reading of two of the 90 negative films, the combined error rate in the example mentioned is about 32%." Garland and others reported high error rates when studies were re-read, all with positive findings. Garland postulated that prospectively reading studies with positive findings, during a routine day, would approach 5%.

Siegle et al.<sup>21</sup> from the University of Texas reviewed the radiology departments of six community hospitals over a seven-year period. Their review included the double-blind review of a sample of 3 to 4% of each radiologist's work at these hospitals. They reviewed more than 11,000 images, including x-rays, fluoroscopy, angiography, ultrasound, CT, mammography, nuclear medicine and magnetic resonance imaging (MRI). The initial interpretations were performed by 35 radiologists belonging to the six different hospitals. The authors found a 2.9% to 5.4% variance in disagreement rates between the six hospitals. The individual radiologist rates of disagreement ranged from 0.8% to 9.2%. The authors reported a 4.4% mean rate of interpretation disagreement for significant errors. They further stated, "Different examination types probably show different degrees of disease prevalence. For example, there is probably a higher incidence of disease depicted on CT scans..." The authors wrote that the 4.4% mean error rate was close to what Garland expected as the error rate of 5% in routine reading of cases.

David Soffa, MD, MPA, FACR and associates from the American Imaging Management, Research Department of the American College of Radiology and Yale University performed a research project on a 26-person radiology group in Dallas, Texas between 1997 and 2001. Two percent of daily cases were double-blind read for quality assurance. The goal of their research was to calculate disagreement rates by radiologist and modality in order to develop a benchmark for use in quality assessment of imaging interpretations.<sup>22</sup> In 6,703 cases submitted for double-blind interpretations read by 26 different radiologists, overall disagreement rate of significant errors was 3.48%. They found an error rate for the following modalities: general radiology (x-ray) 3.03%, mammography 5.79% and ultrasound 4.07%. It is important to note that their study did not include more complicated studies such as CT or MRI examinations. Disagreement rates for

*It is important to define the term "error rate," so as to understand the difference between retrospective analysis of cases with known errors and error rates encountered in the daily interpretation of studies.*

the individual radiologists ranged from 2.04% to 6.90%. Their summary statement was, "A quality assurance program of double reading, involving a 2% random sample of over 300,000 cases produced statistically valid results, yielding a 5% or less disagreement rate between board certified radiologists interpreting plain film, mammography and ultrasound cases."

Abujudeh<sup>23</sup> and associates from the Department of Radiology at Massachusetts General Hospital and Harvard Medical School investigated discrepancy rates for the interpretation of abdominal and pelvic CT examinations among experienced radiologists. Ninety examinations, which were interpreted between May 2006 and April 2007 by one of three designated, Body fellowship-trained expert radiologists with a mean subspecialty radiology experience of 5.7 years, were selected for review. The same radiologists were blinded to the previous interpretations and were asked to re-interpret 60 examinations - 30 of their own previously interpreted cases and 30 interpreted by their colleagues. The interobserver (between two different radiologists) major disagreement rate was 26%; while, the intraobserver (disagreeing with one's self) major disagreement rate was 32%. The work from this prestigious institution confirmed what was suspected by the University of Texas.

Platts-Mills and associates reported in *The Journal of Emergency Medicine*<sup>24</sup> in 2008 a 7% major discrepancy rate for interpretation of abdominal and pelvic CT examinations.

Borgstede et al.<sup>25</sup> reviewed the error rates of 250 radiologists who had interpreted greater than 20,000 examinations, for the clinical testing of RADPEER (American College of Radiology) in 2004. They initially reported a disagreement rate of 3%-3.5%. A subsequent scoring white paper was published in 2009 that reported a RADPEER program discrepancy rate 2.9%.<sup>26</sup> Of concern is the reliance by RADPEER on self-reported data as opposed to other more objective methods of evaluation. As noted by Abujudeh, "Self-reporting of discrepancy could simply underestimate the occurrence as one may elect not to report such discrepancies of their own or those of their peers."

Drs. Zan, Yousem, Carone and Lewin from the Department of Radiology and Radiological Sciences, Johns Hopkins Medical Institution, in their research paper *Second-Opinion Consultations in Neuroradiology*,<sup>27</sup> reported a 7.7% rate of discrepancy interpretations on 4,534 outside- and second-opinion interpretations, further supporting the need for subspecialty interpretations.

*"Self-reporting of discrepancy could simply underestimate the occurrence as one may elect not to report such discrepancies of their own or those of their peers."*

## Discussion

As noted by Garland, "Even experienced physicians are found to have a measurable degree of 'observer error' due apparently to the so-called human equation."

Malcolm Gladwell, writing in *The New Yorker*,<sup>28</sup> made the following observation, "The reason a radiologist is required to assume that the overwhelming number of ambiguous things are normal, in other words, is that the overwhelming number of ambiguous things really are normal."

Radiologists are, in this sense, a lot like baggage screeners at airports. The chances are that the dark mass in the middle of the suitcase isn't a bomb because you've seen a thousand dark masses like it in suitcases before, and none of those were bombs – and if you flag every suitcase with something ambiguous in it, no one would ever make his flight. But that, of course, doesn't mean that it isn't a bomb. All you have to go on is what it looks like on the X-ray screen – and the screen seldom gives you quite enough information.”

The truth is that radiologists are not infallible; and, well-trained, experienced and diligent radiologists will make errors with significant discrepancies. The literature also supports that errors are made by clinicians in all specialties. Re-stated, the unfortunate truth is that radiologists make errors, but do not tabulate, analyze, report, study, and learn from their mistakes.

How and why physicians make errors needs to be studied in depth to further understand what can be done to reduce errors.

Research has provided us with the following information to consider in defining a benchmark in accuracy in radiology:

- Over time, the RADPEER program has reported a discrepancy rate between 2.9% and 3.5%. In this program self-reported error rates are provided to RADPEER as a consensus review of prior images and interpretations, being reviewed for follow-up interpretation. This type of review takes no additional time for a radiology group to do, and is not subjected to a double-blind second interpretation with adjudication by colleagues in an attempt to find the “truth” so its value is questionable. Radiologists inherently under-report error rates. As previously noted by Abujudeh and colleagues,<sup>29</sup> “Self-reporting of discrepancy could simply underestimate the occurrence as one may elect not to report such discrepancies of their own or those of their peers.”
- The University of Texas, in an extensive double blind review, established an error rate 4.4%. They noted that error rates were higher for CT examinations. Rates among individual radiologists varied from 0.8% to 9.2%.
- Soffa and associates in their double-blind review of 2% of 300,000 cases for large groups in Dallas, Texas reported a statistically-valid error rate of 3.48%; however, their study only included x-rays, mammograms and ultrasounds, and did not include CTs or MRIs.
- The Department of Radiology at Massachusetts General Hospital and Harvard Medical School reported on the high variability of fellowship-trained, experienced and dedicated Body radiologists in interpreting CT abdominal and pelvic examinations. The interobserver (disagreement between radiologists) major disagreement rate was 26%. The intraobserver (disagreement with one's self) major disagreement rate was 32%. It is easy to understand how the disagreement between clinicians and radiologists can be equally as extreme for CT of the abdomen and pelvis.

***The truth is that radiologists are not infallible; and, well-trained, experienced and diligent radiologists will make errors with significant discrepancies.***

- Platts-Mills and associates reported in *The Journal of Emergency Medicine*<sup>30</sup> in 2008 a 7% major discrepancy rate for interpretation of abdominal and pelvic CT examinations, interpreted for patients seen through the emergency room.
- The discrepancy rate between general radiologists and subspecialty radiologists can be as high as 7.7%.

## Conclusion

Published research provides us with strong evidence of radiologist error rates in a variety of institutions, settings and modalities. The most accurate determinants of discrepancy rates are provided by double-blind reviews. The blended error rate for a wide range of modalities is 4.4%, with a possible range of errors between radiologists, depending on the type of studies interpreted, modality mix and subspecialty expertise, between 0.8% and 9.2%. General radiologists limiting their interpretations to x-rays, mammograms and ultrasounds will have a collectively lower rate of errors of 3.48%. CT abdominal and pelvis interpretations will have a high disagreement rate between radiologists and clinicians, approaching 32%. The error rate for CT abdominal and pelvic examinations seen from patients through the emergency department will be as high as 7%. The neuroradiology discrepancy rate between general readers and neuroradiologists are estimated to be 7.7%.

### Research and Analysis Contributors

- Radiology Quality Institute
- Radisphere Peer Review Committee
- Radisphere Quality Management Department
- Radisphere Patient Safety Evaluation System Committee

- <sup>1</sup> Pa Patient Saf Advis. 2010 Sep;7(3):76-86.
- <sup>2</sup> Berner ES, Graber ML. Overconfidence as a cause of diagnostic error in medicine. *Am J Med.* 2008 May; 121(5 Suppl):S2-S23.
- <sup>3</sup> Abujudeh, HH, Boland, GW, Kaewalai, R, et al. Abdominal and Pelvic Computed Tomography (CT) Interpretation: discrepancy rates among experienced radiologists. *Eur Radiol.* 2010 Aug;20(8): 1952-1957.
- <sup>4</sup> Garland LH. Studies on the accuracy of diagnostic procedures. *AJR.* 1959;82(1):25-38.
- <sup>5</sup> Siegle, RL, Baram, EM, Stewart, RR, et al. Rates of Disagreement in Imaging Interpretation in a Group of Community Hospitals. *Acad Radiol.* 1998 Mar; 5(3):148-154.
- <sup>6</sup> Berlin, L. Accuracy of Diagnostic Procedures: Has it Improved Over the Past Five Decades? *AJR* 2007;188:1173-1178.
- <sup>7</sup> Siegle, RL, Baram, EM, Stewart, RR, et al. Rates of Disagreement in Imaging Interpretation in a Group of Community Hospitals. *Acad Radiol.* 1998 Mar;5(3):148-54.
- <sup>8</sup> Soffa, DJ, Lewis, RS, Sunshine, JH, et al. Disagreement in Interpretation: A Method for Development of Benchmarks for Quality Assurance in Imaging. *J Am Coll Radiol.* 2004;1(3):212-217.
- <sup>9</sup> Abujudeh, HH, Boland, GW, Kaewalai, R, et al. Abdominal and Pelvic Computed Tomography (CT) Interpretation: discrepancy rates among experienced radiologists. *Eur Radiol.* 2010;20(8): 1952-1957.
- <sup>10</sup> Zan E, Yousem, DM, Carone, M, et al. Second-Opinion Consultations in Neuroradiology. *Radiology.* 2010;255 (1): 125-41.
- <sup>11</sup> Smith MJ. Error and variation in diagnostic radiology. Springfield, IL: C C Thomas, 1967:4, 71, 73, 74, 144-169.
- <sup>12</sup> Stevenson CA. Accuracy of the X-ray report. *JAMA.* 1969;207(6):1140-1141.
- <sup>13</sup> Berlin L. Does the "missed" radiographic diagnosis constitute malpractice? *Radiology* 1977 May;123(2):523-7.
- <sup>14</sup> Markus JB, Somers S, Franic SE, et al. Interobserver variation in the interpretation of abdominal radiographs. *Radiology* 1989;171:69-71.
- <sup>15</sup> Berlin L, Hendrix RW. Perceptual errors and negligence. *AJR* 1998; 170(4):863-867.
- <sup>16</sup> Potchem EJ. Measuring observer performance in chest radiology: some experiences. *J Am Coll Radiol.* 2006 Jun;3(6):423-32.
- <sup>17</sup> Herman PG, Gerson DE, Hessel SJ, et al. Disagreement in chest roentgen interpretations. *Chest* 1975;68 (3):278-82.
- <sup>18</sup> Janjua KJ, Sugrue M, Deane SA. Prospective evaluation of early missed injuries and the role of tertiary trauma survey. *J Trauma.* 1998 Jun;44(6):1000-6.
- <sup>19</sup> Fitzgerald R. Error in Radiology: analysis, standard setting, target instruction and teamworking. 2005 Aug;15(8):1760-7.
- <sup>20</sup> Berlin, L. Accuracy of Diagnostic Procedures: Has it Improved Over the Past Five Decades? *AJR.* 2007;188: 1173–1178.
- <sup>21</sup> Siegle, RL, Baram, EM, Stewart, RR, et al. Rates of Disagreement in Imaging Interpretation in a Group of Community Hospitals. *Acad Radiol.* 1998 Mar;5(3):148-54.
- <sup>22</sup> Soffa DJ, Lew RS, Sunshine JH, et al. Disagreement in interpretation: a method for the development of benchmarks for quality assurance in imaging. *J Am Coll Radiol.* 2004;1(3):212-217.
- <sup>23</sup> Abujudeh, HH, Boland, GW, Kaewalai, R, et al. Abdominal and Pelvic Computed Tomography (CT) Interpretation: discrepancy rates among experienced radiologists. *Eur Radiol.* 2010;20(8): 1952-7.
- <sup>24</sup> Platts-Mills TF, Hendy GW, Ferguson B (2008). Teleradiology interpretations of emergency department computed tomography scans. *J Emerg Med.* HYPERLINK "[http://www.jem-journal.com/article/S0736-4679\(08\)00315-6/abstract](http://www.jem-journal.com/article/S0736-4679(08)00315-6/abstract)"  
abstract "[http://www.jem-journal.com/article/S0736-4679\(08\)00315-6/abstract](http://www.jem-journal.com/article/S0736-4679(08)00315-6/abstract)."
- <sup>25</sup> Borgstede JP, Lewis RS, Bhargavan M, et al: RADPEER quality assurance program; multifacility study of interpretation disagreement rates. *J Am Coll Radiol.* 2004;1(1):59-65.
- <sup>26</sup> Jackson VP, Cushing T, Abujudeh HH et al. RADPEER Scoring White Paper. *J Am Coll Radiol.* 2009;6:21-25.
- <sup>27</sup> Zan E, Yousem DM, Carone M, et al. Second-Opinion Consultations in Neuroradiology. *Radiology.* 2010;255(1):135-41.
- <sup>28</sup> Gladwell M. The picture problem: mammography, air power, and the limits of looking. *The New Yorker.* December 13, 2004:74-81.
- <sup>29</sup> Abujudeh HH, Boland GW, Kaewalai R, et al. Abdominal and Pelvic Computed Tomography (CT) Interpretation: discrepancy rates among experienced radiologists. *Eur Radiol.* 2010;20(8): 1952-7.
- <sup>30</sup> Platts-Mills TF, Hendy GW, Ferguson B (2008). Teleradiology interpretations of emergency department computed tomography scans. *J Emerg Med.* 2010;38(2):188-195.