Methods toward improving consistency in interpretation of chest radiographs

Presented by:
Dr. D.A. Soboleski
Associate Professor
Queen’s University
Kingston, Ontario
Numerous studies have documented the poor inter/intra-observer agreement on interpretation of chest radiographs.
Reliability in interpretation is crucial to guide appropriate timely care particularly in the very sick population

An effect of inconsistent interpretation often overlooked

Many research study protocols rely on the radiographic interpretation to determine which arm of the study a patient may proceed upon.
Reliability helps ensure reproducibility in clinical research studies allowing for a reduced sample size requirements and allow true-positive findings.
Why the poor inter/intra-observer agreement?

Is it due to our differences in perception?

Science is nothing but perception.

— Plato —
Definition:

Perception is the process of selecting and interpreting the information we receive thru our senses to produce a meaning and plays a significant role in the interpretation of images.
How can we improve perception and consistency?

SOLUTIONS
1. Ensure **we and our colleagues** are looking at the same image.
Effect of PACS image manipulation on the agreement of chest radiograph interpretation in the NICU

Collaborators

D. Castro - pediatric radiologist
M. Flavin – neonatologist
M. Clarke – neonatologist
J. Flood – thoracic radiologist
J. Gammon – neonatologist/pediatrician
Prospective cohort study - the population

- 60 patients
- gestational age 26-32 weeks
- 1 day – 3 months of age
- all with history of surfactant deficiency disease
Prospective cohort study

- 2 chest x-rays on each pt. performed on different days included

- the 120 x-rays (60 pts.x 2) anonymized and numbered

- randomly placed as acquired in two identical viewers
Prospective cohort study

- 3 radiologists - 3 neonatologists

Radiologists – median length of expertise 14 yr (5-25)

Neonatologists – median length of expertise 16 yr (2-25)
Prospective cohort study

- Reviewed 2 consecutive x-rays on same pt.
  - Once in usual manner with ‘windowing’/image manipulation allowed
  - Once again without image manipulation
Effect of PACS image manipulation on the agreement of chest radiograph interpretation in the NICU

only interpretation

is disease appearance

• better
• worse
• unchanged
**Results**

Assessment of sixty sets of frontal chest radiographs ‘without’ versus ‘with’ the ability to manually manipulate the images

<table>
<thead>
<tr>
<th></th>
<th>Nonmanipulated (n=60) (%)</th>
<th>Manipulated (n=60) (%)</th>
<th>Relative risk (95% CI)</th>
<th>$\chi^2$ (P)</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Perfect agreement</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Combined (n=6)</td>
<td>8 (13)</td>
<td>3 (5.0)</td>
<td>2.9</td>
<td>2.50 (0.11)</td>
</tr>
<tr>
<td>Radiologists (n=3)</td>
<td>25 (42)</td>
<td>21 (35)</td>
<td>1.3</td>
<td>0.56 (0.46)</td>
</tr>
<tr>
<td>Neonatologists (n=3)</td>
<td>18 (30)</td>
<td>8 (13)</td>
<td>2.8</td>
<td>4.91 (0.04)</td>
</tr>
</tbody>
</table>
## Results

Assessment of sixty sets of frontal chest radiographs ‘without’ versus ‘with’ the ability to manually manipulate the images

<table>
<thead>
<tr>
<th></th>
<th>Nonmanipulated (n=60) (%)</th>
<th>Manipulated (n=60) (%)</th>
<th>Relative risk (95% CI)</th>
<th>( \chi^2 ) (P)</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Opposing agreement</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Combined (n=6)</td>
<td>29 (48)</td>
<td>37 (62)</td>
<td>0.6</td>
<td>2.16 (0.14)</td>
</tr>
<tr>
<td>Radiologists (n=3)</td>
<td>12 (20)</td>
<td>17 (28)</td>
<td>0.6</td>
<td>1.12 (0.29)</td>
</tr>
<tr>
<td>Neonatologists (n=3)</td>
<td>11 (18)</td>
<td>24 (40)</td>
<td>0.3</td>
<td>6.82 &lt;0.01</td>
</tr>
</tbody>
</table>
How can we improve perception and our consistency in reporting?

SOLUTION

1. Do not allow image manipulation outside imaging the imaging department
How can we improve perception and consistency?

SOLUTIONS

1) Ensure we and our colleagues are looking at the same image

2) Ensure we are looking at the same image
VALUE OF A NOVEL DEVICE AND METHOD ALLOWING FOR IMAGE EQUALIZATION AND SYNCHRONIZATION OF MANUAL WINDOWING, WHEN COMPARING RECENT CHEST RADIOGRAPHS WITH PREVIOUS STUDIES

Collaborators

Denise Castro, pediatric radiologist
S.Salahudeen, thoracic radiologist (asst.prof)
R.Nolan, thoracic radiologist (prof)
E.VanDenKerkhof, epidemiologist, PhD (prof)
Dr.Flood, thoracic radiologist
Purpose

to determine whether a novel DEVICE and METHOD, that equalizes chest radiographic appearance and allows for synchronization of manual windowing with comparison studies, would improve consistency in interpretation and dictating efficiency.
What prompted the study?

Previous studies have demonstrated a sub-optimal chest radiograph appearance in up to one-third of cases and have shown a poor correlation with autopsy findings.
We were spending way too much time and effort agonizing over potential changes on the radiograph.
It can be extremely difficult for the technicians to reproduce exact positioning and exposure techniques particularly in the ICU/NICU. Patient habitus may also change between studies, i.e., weight gain or loss, surgery, etc.

Support apparatus may change exposure as well as prevent proper positioning. Shorter beam distances than recommended in some pts all contribute to that 1/3 of cases which are suboptimal.
The clinical question often breaks down to

- *is it better or worse*
The Novel Device

- consists of a VAP (variable attenuation plate) composed of variable thicknesses of different metals, i.e., brass, Al

0.1 inch brass mounted on 0.0625 inch aluminum
The Novel Device

- plate can be varying shapes, with or without ‘handle’

2cm by 2 cm square or 1 cm by 4 cm strip

each square a different thickness based on no. of layers
The Novel Device

- upon exposure 4 differing density quadrants
The Novel Method consisted of software specifically designed to ‘match’ each quadrant on the VAP with its corresponding quadrant on a separate image.
An high value of correlation (close to 1) indicates that the intensities are distributed linearly, therefore validating the methodology because the calibration of the images is independent on which intensity have been used.
Locating the Markers
Synchronized Window Level
Study methods

- Conducted at an ICU in a tertiary care hospital
- 50 non-consecutive patients
- 17 days to 85 yrs of age (24 males)
- 29/50 intubated
Study methods

• The 50 patients undergoing CXRs as part of their routine care

• Each patient had frontal radiographs (on different days) in which the VAP had been placed on the cassette

• Important - the VAP needed to be away from the patient and support apparatus
Study methods

- Short training session to 5 technical staff
  - explain importance of positioning of VAP
  - instructed to place similar to their rt \ lt markers
  - technique otherwise unaltered
    90kVp and 4mA (60kVp and 1.5mA nicu)
Study Method

- the 100 radiographs (50pts x 2 CXRs) were then randomly placed as acquired, in 2 identical viewers on our PACS, with the previous study to the right of the more recent CXR.

Cases were anonymized and identified by number.

3 thoracic radiologists then reviewed and interpreted the 100 cases (50 paired cases in each viewer).

Experience 3 – 30 years- no conflicting commercial interest.
Study Method

- dictation of the 50 cases was conducted in the radiologists usual manner using one viewer with windowing as deemed appropriate

- on the other viewer, dictation was conducted with use of the specifically designed software utilizing the VAP allowing for equalization of appearance and synchronization of windowing when comparing recent to previous CXRs

- sequence of dictation was randomly alternated between those utilizing the VAP and those not
Study Method

- each report included an impression of either
  - WORSE
  - NO CHANGE
  - IMPROVED

- Dictation time per case and total was calculated by an observer

- Report impressions were compared both between radiologists and individually between methods (with and without use of the VAP/software)
<table>
<thead>
<tr>
<th>CASE</th>
<th>Radiologist A</th>
<th>Radiologist B</th>
<th>Radiologist C</th>
</tr>
</thead>
<tbody>
<tr>
<td>#1</td>
<td>X X</td>
<td>X</td>
<td>X</td>
</tr>
<tr>
<td>#2</td>
<td>O O</td>
<td>O</td>
<td>O</td>
</tr>
<tr>
<td>#3</td>
<td>O O</td>
<td>O</td>
<td>O</td>
</tr>
<tr>
<td>#4</td>
<td>X X</td>
<td>X</td>
<td>X</td>
</tr>
<tr>
<td>#5</td>
<td>😊</td>
<td>O</td>
<td>😊</td>
</tr>
<tr>
<td>#6</td>
<td>X O</td>
<td>O</td>
<td>X</td>
</tr>
<tr>
<td>#7</td>
<td>X O</td>
<td>O</td>
<td>O</td>
</tr>
<tr>
<td>#8</td>
<td>X 😊</td>
<td>😊</td>
<td>😊</td>
</tr>
<tr>
<td>#9</td>
<td>O X</td>
<td>O</td>
<td>X</td>
</tr>
<tr>
<td>#10</td>
<td>O O</td>
<td>O</td>
<td>O</td>
</tr>
</tbody>
</table>

X = worse  O = no change  😊 = improved
Results

• Each radiologist dictated all 50 cases in each viewer. (One case discarded due to dictation problem)

• There was a statistically significantly difference in agreement on case impression between the two methods
Results

- Kappa values between Rads A and B, A and C, B and C

  without VAP  46%  55%  51%

  with VAP    73%  81%  66%
Results

- Intra-observer weighted Kappa values between non-VAP and VAP methods
  - ranged 63% - 86%
  - suggested moderate to very good agreement between methods
Results

Discrepant report impressions

• 1 rad stating improved vs one or both other rads stating unchanged or worse

• without VAP range of 24 – 28 %

• with VAP range of 10 – 16% (p<0.01).
Opposing report impressions

- 1 rad stating improved vs one or both other rads stating worse
- without VAP - 12 % of cases
- with VAP - 7 % of cases

Results
Results

• the mean time to dictate each case was 44 seconds for both groups

• however, the mean total time to dictate the whole batch of radiographs was 20 minutes faster (97 min. vs 77 min.) using the VAP method

(includes the actually dictation ‘mic’ time as well as the time the radiologist spent deciding what to say in their report)
Conclusion

• **IMPROVED CONSISTENCY** in report impressions which we believe can result in improved patient care

• **EFFICIENCY** - the novel method allowed for an ~20% decrease in overall reporting time in our study
How can we improve perception and our consistency in reporting?

SOLUTION

2. Equalize the appearance and allow manual window synchronization of recent CXRs with their previous study
Why the poor inter-observer agreement?

Problem solved?

If perceptual? Not just technical solutions

inherent
How many faces can you perceive?
Definition of diagnostic error

Diagnostic error is defined as a diagnosis which is missed, delayed or wrong as determined by a subsequent definitive exam or test.

interpretive vs perception

Perception error up to 80 %
First described by Garland in 1949
• perception error in 20-30% of chest cases in a study on TB

More recent perception error studies
• across multiple modalities
• across multiple countries and
• across multiple subspecialties

NOT changed significantly since Garland first described
A medical image perception society (MIPS) now exists with a goal to improve the understanding of imaging perceptual factors and foster research.
attempts to improve perception/decrease perception error and thus improve consistency in interpretation of all imaging modalities

- structured reporting
- Double/triple reading
- Improved luminance
- Changing focal zones

Gaze trackers
- limited success so far
Factors affecting perception error in Sonography

Collaborators

D. Castro – pediatric radiologist
E. Sauerbrei – radiologist
M. Kolar – pediatric surgeon
W. Hopman - biostatician
Purpose

- determine if experience
- knowledge base
- amount of time spent

correlated with ability to see a normal appendix
Study population

- 343 pts referred to imaging dept. with a clinical concern of appendicitis

- pts with a normal appendix or non-visualized appendix with subsequent discharge and normal follow-up included
Sonograms performed in the usual manner by

- one of 15 trained sonographers
  (3-23 yrs experience)

- one of 8 radiology residents
  (4 in PGY 2/3; 4 in PGY 4/5)
Result – overall adult population

• no effect knowledge base
• no effect experience level
• no effect time spent
If patient could have been triaged to the 4 sonographers with the best perception in identifying the normal appendix, the success rate would increase to 75% from 27%. This would have resulted in a 48% decrease in the number of CT’s ordered to further evaluate these patients.
Take home message

1. new and novel methods are needed to improve our consistency in interpretation of imaging studies

2. despite our advancements perception error will always play a role in our daily lives
Thank-you for your attention

The Bittersweet Meeting of Perception and Reality