Weighted kappa as an alternative method to simple kappa statistics: evaluating inter-reader reliability in ILO review of chest radiographs

Jacob Buhrow, MD/MPH Candidate
Mentor: Dr. Laurence Fuortes

Background: ILO Classification System

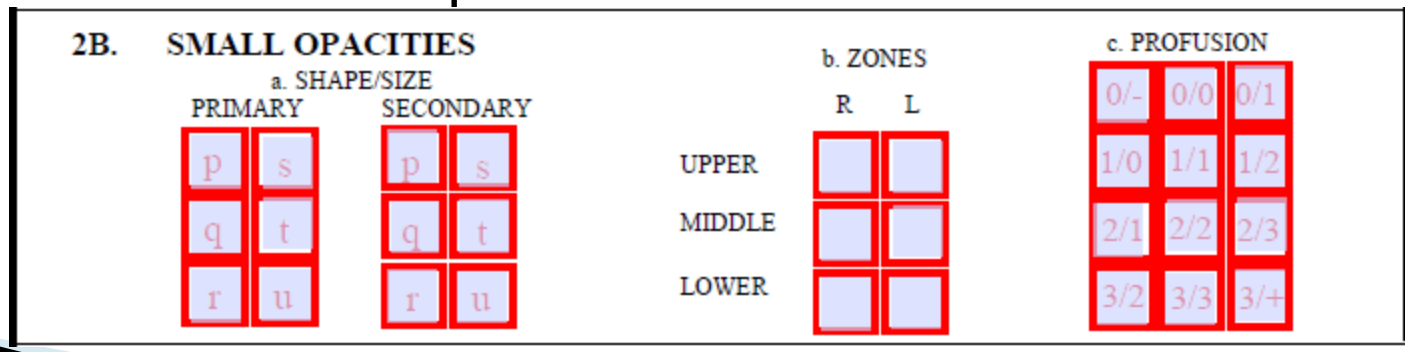
International Labor Office (ILO) Classification of Radiographs of Pneumoconioses

- Epidemiologic classification
- Created in 1930
- Revised in 1950, 1958, 1968, 1971, 1980, 2000
- ILO standard classification form

Background: ILO Classification System

Profusion of small opacities

- Determined by comparison of radiograph with standardized reference films
- Reference films represent midrange of four major ordinal categories (0–3)
- Minor categories
  - Help reader place radiograph on continuum
  - Create twelve-point scale



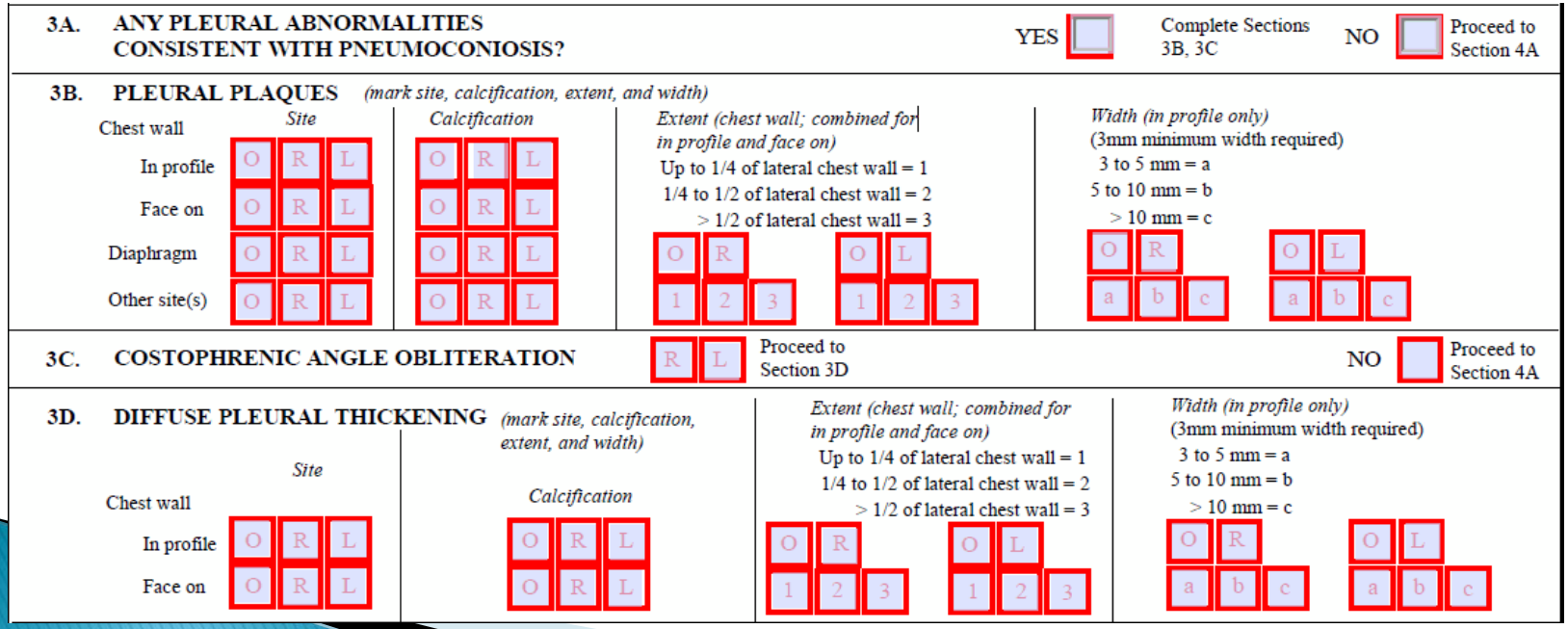
NIOSH Topic: Chest Radiography: B Reader Information for Medical Professionals | CDC/NIOSH." *Centers for Disease Control and Prevention*. 08 Jan. 2010. Web. 10 Sept. 2010. <http://www.cdc.gov/niosh/topics/chestradiography/breader-info.html#e


s/s – 1/1
s/s – 2/2
normal – 0/0
s/s – 3/3
NIOSH SS - 006

Background: ILO Classification System

Pleural abnormalities

- Recorded as yes/no and include:
  - Pleural Plaques
  - Costophrenic angle obliteration
  - Diffuse pleural thickening



3A. ANY PLEURAL ABNORMALITIES CONSISTENT WITH PNEUMOCONIOSIS? YES ☐ Complete Sections 3B, 3C NO ☐ Proceed to Section 4A

3B. PLEURAL PLAQUES *(mark site, calcification, extent, and width)*

Chest wall	Site	Calcification	Extent (chest wall; combined for in profile and face on)	Width (in profile only) (3mm minimum width required)
In profile	O R L	O R L	O R 1 2 3 / O L 1 2 3	O R a b c / O L a b c
Face on	O R L	O R L		
Diaphragm	O R L	O R L		
Other site(s)	O R L	O R L		

Up to 1/4 of lateral chest wall = 1; 1/4 to 1/2 of lateral chest wall = 2; > 1/2 of lateral chest wall = 3

3 to 5 mm = a; 5 to 10 mm = b; > 10 mm = c

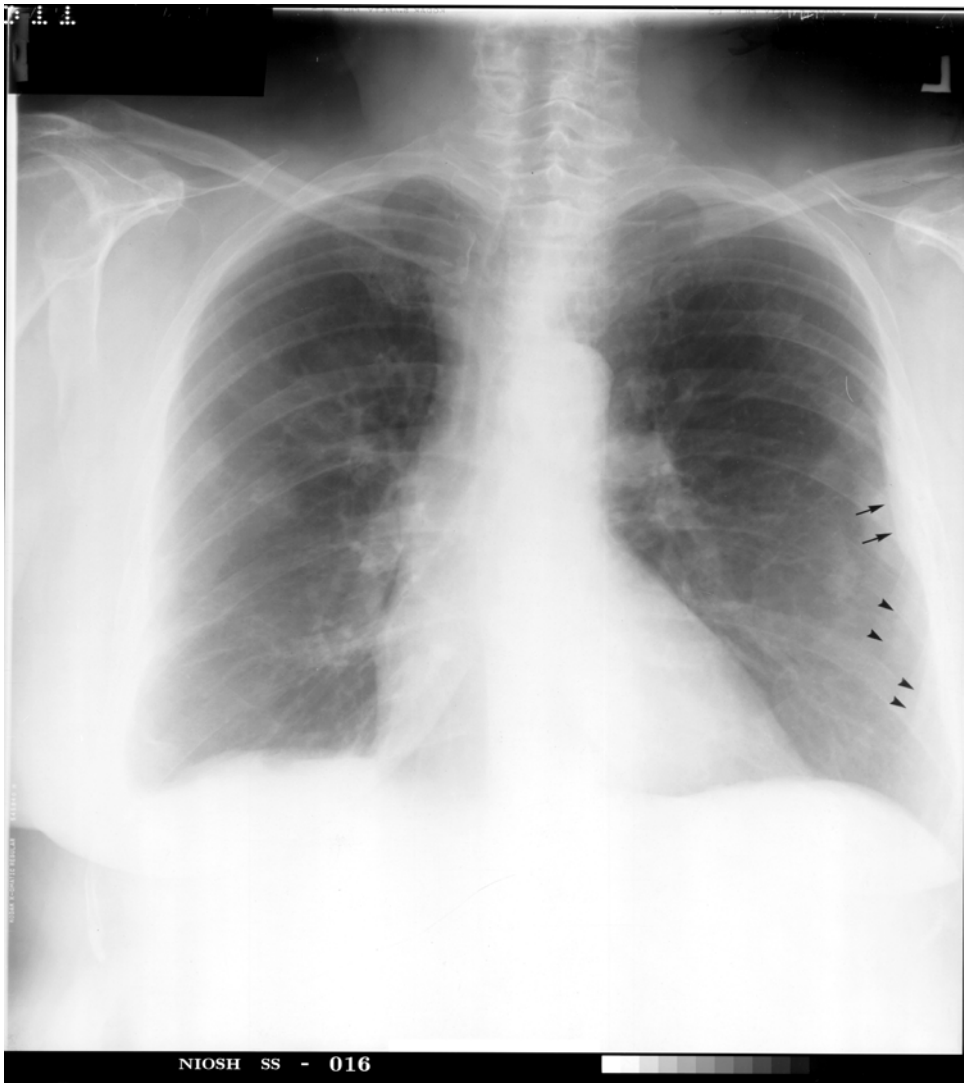
3C. COSTOPHRENIC ANGLE OBLITERATION R L Proceed to Section 3D NO ☐ Proceed to Section 4A

3D. DIFFUSE PLEURAL THICKENING *(mark site, calcification, extent, and width)*

Chest wall	Site	Calcification	Extent (chest wall; combined for in profile and face on)	Width (in profile only) (3mm minimum width required)
In profile	O R L	O R L	O R 1 2 3 / O L 1 2 3	O R a b c / O L a b c
Face on	O R L	O R L		

Up to 1/4 of lateral chest wall = 1; 1/4 to 1/2 of lateral chest wall = 2; > 1/2 of lateral chest wall = 3

3 to 5 mm = a; 5 to 10 mm = b; > 10 mm = c


NIOSH SS - 016

Background: ILO Cohort

Iowa Army Ammunition Plant

- Located in Burlington, IA
- Employed thousands of people in the production of traditional (1941–present day) and nuclear (1949–1975) weapons
- Atomic Energy Commission (AEC)

Ames Lab

- Contracted by US government to work on Manhattan Project

Background: ILO Cohort

Gender	Male: 582	Female: 197	
Age at time of X-ray	High: 99	Low: 25	Average: 68
Ethnicity	Caucasian: 733 Asian American: 3	African American: 12 Native American: 2	Hispanic American: 8 No Data: 21
Smoking History	Ever Smoker: 422	Never Smoker: 350	No Data: 7

Cohen's Kappa Statistic: Overview

- Measurement of inter-rater agreement
  - Provides quantitative estimate of reliability
  - Predicts precision *not* accuracy
  - Adjusts for chance agreement
    - Values fall between -1 and 1
      - "-1" = complete disagreement
      - "0" = agreement expected by chance
      - "1" = perfect agreement
- Simple Kappa
  - Credit given only for complete agreement
- Weighted Kappa
  - Credit given for complete and partial agreement

Cohen J. A coefficient of agreement for nominal scales. Educational and Psychological Measurement 1960;20:37-46.

Methods

Sample selection

- Over 2000 individuals from cohort had chest X-ray as part of Former Worker Program screening
  - Three physicians experienced in ILO rating reviewed X-rays

Inclusion criteria

- ILO form completed by all three physicians
- X-ray of sufficient quality

Methods

Final data set

- Included blinded ILO data for 779 chest X-rays
- SAS 9.1.3 statistical software used to calculate agreement between pairs of physicians (1,2; 1,3; 2,3) with regard to four specific data sets/groupings
  - Ungrouped profusion ratings (0/0, 0/1, 1/0, 1/1, 1/2, 2/1, 2/2, 2/3, 3/2)
  - Grouped profusion ratings (0/0, 0/1 = 0; 1/0, 1/1, 1/2 = 1; 2/1, 2/2, 2/3 = 2; 3/2 = 3)
  - Parenchymal abnormalities only (Y or N)
  - Pleural abnormalities only (Y or N)

Results

	Number of Ratings (Percent Distribution of Ratings)		
Rater Profusion category	1	2	3
0/0	708 (89.39)	678 (85.61)	647 (81.69)
0/1	33 (4.17)	44 (5.56)	69 (8.71)
1/0	27 (3.41)	28 (3.54)	45 (5.68)
1/1	20 (2.52)	19 (2.40)	17 (2.15)
1/2	1 (0.13)	16 (2.02)	9 (1.14)
2/1	1 (0.13)	6 (0.76)	2 (0.25)
2/2	2 (0.25)	0 (0.00)	2 (0.25)
2/3	0 (0.00)	0 (0.00)	1 (0.13)
3/2	0 (0.00)	1 (0.13)	0 (0.00)

Results

Raters	1 / 2	1 / 3	2 / 3
Simple kappa: Ungrouped profusion	0.37	0.29	0.35
Weighted kappa: Ungrouped profusion	0.48	0.47	0.55
Simple kappa: Grouped profusion	0.45	0.50	0.58
Simple kappa: Parenchymal Y or N	0.53	0.56	0.66
Simple kappa: Pleural Y or N	0.75	0.68	0.79

All values $p<0.0001$

Results

Kappa values

- Represent agreement greater than that expected by chance alone
- Fall within the range of fair to substantial according to suggestions of Landis and Koch

Kappa	Agreement
0.01–0.20	Slight
0.21–0.40	Fair
0.41–0.60	Moderate
0.61–0.80	Substantial
0.81–0.99	Almost perfect

Landis JR, Koch GG. The measurement of observer agreement for categorical data. Biometrics 1977;33:159–74.

Questions?