

Evaluating evidence of assessor disagreement
[was *A Randomization Test of Rater Agreement with Groups of Raters*]

A. John Bailer^{1,2} and Robert B. Noble¹

¹Department of Mathematics & Statistics

²Scripps Gerontology Center

Miami University

Oxford, Ohio 45056

T: 513.529.3538

F: 513.529.1493

Email: baileraj@muohio.edu

Other Collaborators: Jane Straker, Mike Hughes

ENAR 2007– Session 58. CONTRIBUTED PAPERS: HEALTH SERVICES RESEARCH AND MEDICAL COST – Tuesday March 13, 8:30-8:45

Outline:

1. Background
2. Study and data structure
3. If gaming, then what is expected?
4. So how can we test this?
5. What did we see?
6. Afterthoughts

1. Background

- * The health status and associated resource utilization for every nursing home resident is evaluated quarterly [and upon some change of status (e.g. readmission to the nursing home after a stay in a hospital)]
- * Minimum Data Set (MDS) = standard assessment instrument
- * MDS usually completed by a RN ("MDS or resident assessment coordinator") and includes input from other professionals and staff working at the nursing home (e.g. nutritionist, occupational therapist, etc.).

- * MDS assessment yields a Resource Utilization Group (RUGS) category for each resident.
- * Each level in this **44 level categorization** is then assigned a numeric score, the **case-mix score** \approx amount of skilled nursing care needed by a resident.
- * Case-mix scores are used to help set reimbursement levels for a facility so that facilities with higher case-mix scores receive higher levels of reimbursement.
- * **MOTIVATION:** one might worry that facilities might feel pressured to maximize their case-mix scores in order to generate the largest reimbursement.

2. Study and data structure

- * we conducted a study in which independent assessors of an indicator of resource utilization, case-mix scores, were compared to nursing-home assessors.
- * an independent team of Registered Nurse assessors were sent out to a sample of nursing homes in Ohio.
- * case-mix scores derived by this team of independent assessors for a sample of residents in each facility was compared to the case-mix scores generated by the nursing homes for the same residents.

Assessor (r)	# Facilities = m_r (residents/facility)	NH rating, indep. rating	A<F	A=F	A>F
1 ("K")	5 (2-4, 9)				
2 ("Sm")	5 (1, 3-7)		...		
3 ("Sp")	5 (4-10)		...		
4 ("T")	7 (1, 8-10)		...		
5	13		...		

("W")	(2-10)		
6	4		...
("Z")	(3-7)		

For a given independent assessor going to a particular nursing home.

$A < F$	$A = F$	$A > F$
n_{1ij}	n_{2ij}	n_{3ij}

where $1 \leq n_{1ij} + n_{2ij} + n_{3ij} \leq 10$ for $i=1, \dots, 6$ and $j=1, \dots, n_i$

3. If gaming, then what is expected?

For a given independent assessor going to a particular nursing home.

$A < F$	$A = F$	$A > F$
π_1	π_2	π_3

$$\pi_1 + \pi_2 + \pi_3 = 1$$

If Facility Assessor systematically rates residents Higher than the Independent assessor, then $\pi_1 > \pi_3$.

4. So how can we test this?

4.1 Are the independent assessors the same?

4.2 What would you expect for each assessor?

4.3 Which facilities are flagged?

4.1 Are the independent assessors the same?

- * permutation test (see, e.g., Good 2000) - facility labels of residents at the 39 facilities were randomly permuted.

- * 5000 permutations of the data

- * disagreement % was then calculated for each permuted data set

- * observed disagreement percentage was compared to this permutation distribution and a permutation P-value was obtained

4.2 What would you expect for each assessor?

Under H_0 : same probability of independent assessor (A) generating a lower or higher rating than the facility assessor (F),

A<F	A=F	A>F
π	$1-2\pi$	π

which for observed data

A<F	A=F	A>F
n_1	n_2	n_3

$$\text{would yield } \hat{\pi} = \frac{n_1 + n_3}{2 \times (n_1 + n_2 + n_3)}$$

4.3 Which facilities are flagged?

- i. calculate $\hat{\pi}$ for each assessor (assuming heterogeneity in the independent assessors)
- ii. calculate the $\Pr(N_1=n_1, N_2=n_2, N_3=n_3 \mid \hat{\pi})$ for the observed data
- iii. P-value = $\sum \Pr(N_1=n_1, N_2=n_2, N_3=n_3 \mid \hat{\pi})$ where the sum is taken over all configurations of counts equal to or more extreme than observed in the direction of $A < F$.

For example, Assessor "Sp" had $\hat{\pi}=0.314285714$. Thus, we consider any set of resident ratings at a facility to be the realization of a multinomial random variable with $\hat{\pi}=(0.314, 0.371, 0.314)$

Consider the P-value for a particular provider where $n=10$ residents were assessed where 6 cases with the assessor higher than the assessor ($A<F$), 1 tied case ($A=F$), and 3 cases where the assessor rated the resident higher than the facility ($A>F$).

A<F	A=F	A>F
6	1	3

More extreme would be cases shift towards "A<F", e.g.

A<F	A=F	A>F
6	2	2

The possible values as extreme, or even more so, than observed are

A<F	A=F	A>F	Probability
6	4	0	0.003851841
7	3	0	0.001862429
8	2	0	0.000590963
9	1	0	0.000111121
10	0	0	9.40257E-06
6	3	1	0.013037001
7	2	1	0.004727704
8	1	1	0.001000091
9	0	1	9.40257E-05
6	2	2	0.016546963
7	1	2	0.004000365
8	0	2	0.000423115

6	1	3	0.009334184
7	0	3	0.001128308

The probability for the first case (for example) is

$$\frac{10!}{6!4!0!} 0.314285714^6 0.371428571^4 0.314285714^0 = 0.003851841$$

The p-value (0.0567) is the sum of the above listed probabilities

5. What did we see?

5.1 Are the independent assessors the same?

Nope.

Assessors ranged from 22% to 90% agreement with the facility assessors.

A permutation test of Independent Assessor homogeneity yielded a P-value < 0.01 for 3 of the independent assessors.

5.2 What would you expect for each assessor?

Assessor	# Facilities	A<F	A=F	A>F	$\hat{\pi}$	% agree
K	5	5	14	1	0.15	70
Sm	5	2	14	2	0.11	78
Sp	5	11	14	11	0.31	38
T	7	19	29	3	0.22	56
W	13	2	71	6	0.05	90
Z	4	8	4	6	0.39	22

5.3 Which facilities are flagged?

- * P-value < 0.10 for 8 of 39 facilities
- * 2 Assessors did not flag any facilities (5 and 4 facilities, resp.)
- * 2 Assessors flagged 1 facility (1 of 5, 1 of 13)
- * 1 Assessor flagged 2 facilities (2 of 5)
- * 1 Assessor flagged 4 facilities (4 of 7)

6. Summary

- * Employed a multinomial prob. calculation as the basis of detecting disagreements between an independent assessor and a facility assessor

- * differences could be attributed to differences between ...

1. the facility and the independent assessor

2. independent assessors

3. facility assessors

[different data needed to tease this out ...]

- * can't separate an outlying independent assessor from a collection of nursing facilities that are systematically overestimating case-mix scores
- * look at this as a screening tool to suggest further investigation of facilities