

Improving Data Quality

Using Inter-rater Reliability (IRR)¹

Laura Cassidy, PhD

¹Cassidy LD, Marsh GM, Holleran MK, Ruhl LS,
Methodology to improve data quality from chart review in the managed care setting.
Am J Manag Care. 2002 Sep;8(9):787-93.

Kappa Statistic

- To measure how observers classify individual subjects into the same category on the measurement scale
- A chance-corrected measure of agreement between raters

Data Layout for Kappa Statistic

- **a** = Both Doctors diagnosed disease +
- **b** = Doctor A disease +, Doctor B disease -
- **c** = Doctor A disease -, Doctor B disease +
- **d** = Both Doctors diagnosed no disease -

		Doctor B		
		Disease	No Disease	Total
Doctor A	Disease	a	b	a + b
	No Disease	c	d	c + d
	Total	a + c	b + d	n

Calculation of Kappa Statistic

p_0 = observed proportion of agreement = $(a+d) / n$

p_e = expected proportion of agreement

$p_e = (a + c)(a + b)/n^2 + (b + d)(c + d)/n^2$

$\kappa = (p_0 - p_e) / (1 - p_e)$

		Doctor B		
		Disease	No Disease	Total
Doctor A	Disease	a	b	a + b
	No Disease	c	d	c + d
	Total	a + c	b + d	n

Relationship of Observed to Expected

		Doctor B		
		+	-	Total
Doctor A	+	a	b	a + b
	-	c	d	c + d
	Total	a + c	b + d	n

$$\kappa = (p_0 - p_e) / (1 - p_e)$$

If $(b=c=0)$ perfect agreement $\longrightarrow \kappa = 1.0$

If $p_0 \geq p_e$ $\longrightarrow \kappa \geq 0$

If $p_0 \leq p_e$ $\longrightarrow \kappa \leq 0$

If $p_e = 0.5$ $\longrightarrow \kappa = -1.0$

Interpretation

K

Strength of Agreement

0.75-1.0

Excellent reproducibility

0.4-0.74

Good reproducibility

<0.39

Poor reproducibility

Example

- Chart review is an integral part of data collection in managed care
- Bias may exist due to multiple reviewers
- Even well designed data collection tools are at risk of miss-classification
- Recommendations based on data from chart review are only as reliable as the data collected

Objective

- We developed a standardized approach to inter-rater reliability (IRR) methods onsite/medical record reviews to:
 - Determine extent of agreement between multiple reviewers
 - Identify areas for improvement of data collection procedures
 - Improve data reliability

Methods

- Statistically valid sampling based on individual HEDIS Hybrid measures
- Random sample per measure is evaluated by standard nurse reviewer and an in-house “gold standard”

Example

$$p_0 = (26+44) / 100 = 0.70$$

$$p_e = (40 \times 42) / 100^2 + (58 \times 60) / 100^2 = 0.516$$

$$\kappa = (0.70 - 0.516) / (1 - 0.516) = 0.375$$

		Nurse 2		
		Met Criteria	Did not meet	Total
Gold Standard Nurse	Met Criteria	26	14	40
	Did not meet	16	44	60
	Total	42	58	100

Process Improvements

- Revise data collection tool to reduce subjectivity
- Annual retraining focused on issues identified in previous years
- Measures identified for improvement resulted in excellent IRR in subsequent years

Results

- Across all years most measures showed excellent IRR (0.75-1.0).
- The following showed room for improvement:
 - 1997: PAP (k=0.50)
 - 1998: Well child visits 3-6 yrs (k=0.37)
 - 1999: Comprehensive diabetes (k=0.73)
 - 1999: High blood pressure (k=0.73)
 - 2000: Well child 1st 15 months (k=0.48)

Sample Size Calculations

- $N = Z_{\alpha}^2 p_0 (1 - p_0) \delta^2 / \delta^2 (1 - p_e)^2$
- N = required sample size
- $Z_{\alpha} = 1.96$
- δ = 1/2 the width of the confidence interval (generally 0.1 or 0.2)

Sample Size

$$N = Z_{\alpha}^2 p_0(1 - p_0) / \delta^2(1 - p_e)^2$$

$$N = (1.96)^2(0.70)(1-0.70) / (0.1)^2(1-0.516)^2$$

$$N = 344$$

For lower precision and smaller sample size use $\delta^2 = 0.2^2$ (90% CI)

Ongoing Enhancements

- Utilize in-house nurses each year and annually undergo training
- Eliminated paper abstraction tools
- Laptops standardize data collection activities and improves data reliability
- Blinded IRR record selection reduces bias and promotes data collection efficiency
- Continue using statistically sound IRR study to ensure data reliability

Conclusion

- HEDIS[®] measures are evaluated annually, therefore, regular IRR studies offer ongoing opportunity for process improvements
- High kappa values provide confidence in data reliability and subsequent conclusions
 - Low kappa values identify areas for improvement

References

- **Measurement of the kappa value involving more than two observers; see Fleiss JL, Cuzick J. The reliability of dichotomous judgements: unequal numbers of judgements per subject. *Applied Psychological Measurement* 1979;3:537-42.**
- **Calculation of weighted kappa; see Dunn G. *Design and analysis of reliability studies: Statistical evaluation of measurement errors*. Edward Arnold 1994.**
- **Dunn G, Everitt B. *Clinical Biostatistics: An Introduction to Evidence-Based Medicine*. London: Edward Arnold 1995.**
- **Everitt BS. *Statistical Methods for Medical Investigations*, 2nd ed. London: Edward Arnold 1994.**
- **Calculator <http://www.dmi.columbia.edu/homepages/chuangj/kappa/>**