Technical Brief ■

Agreement, the F-Measure, and Reliability in Information Retrieval

George Hripcsak, MD, MS, Adam S. Rothschild, MD

Abstract Information retrieval studies that involve searching the Internet or marking phrases usually lack a well-defined number of negative cases. This prevents the use of traditional interrater reliability metrics like the κ statistic to assess the quality of expert-generated gold standards. Such studies often quantify system performance as precision, recall, and F-measure, or as agreement. It can be shown that the average F-measure among pairs of experts is numerically identical to the average positive specific agreement among experts and that κ approaches these measures as the number of negative cases grows large. Positive specific agreement—or the equivalent F-measure—may be an appropriate way to quantify interrater reliability and therefore to assess the reliability of a gold standard in these studies.

■ **J Am Med Inform Assoc.** 2005;12:296–298. DOI 10.1197/jamia.M1733.

Informatics researchers frequently perform classification studies in which the number of cases is not definite. Define a *positive case* as a case with some attribute of interest. For example, if cases are documents, then a positive case might be a document that is relevant to some user query. Define a *negative case* as a case that lacks that attribute. It is frequently found that the number of negative cases is uncountable or undefined. In studies of information retrieval[1] of Internet documents, computer systems select relevant documents from the Internet. Negative cases correspond to all nonrelevant Internet documents. Their number is very large, poorly defined, and constantly changing. In text markup studies, computer systems mark relevant phrases in documents. Negative cases correspond to nonrelevant phrases. Their number is poorly defined because phrases can overlap and vary in length.

In many of these studies, there is no perfect gold standard. Instead, researchers rely on experts to generate a gold standard.[2] The experts carry out the same classification tasks as the computer system that is being studied, and the experts' answers are aggregated (for example, using majority opinion) into a single set of best answers that serve as a gold standard. To assess the computer system's performance, the system's answers are compared with the aggregated answers of the experts.

The researcher must assess the quality of this expert-derived gold standard before comparing any system with it. In measurement theory, *reliability* quantifies the degree to which a measurement is repeatable[3]; an unreliable measurement is noisy. A good gold standard must be reliable. Researchers usually quantify reliability using reliability coefficients or κ statistics.[4] When experts create a gold standard in a classification experiment, the reliability of the gold standard can be expressed as the *interrater agreement*, which is the agreement among the experts. This is usually quantified by the κ statistic, which is the chance-corrected interrater agreement (that is, agreement among the experts above that expected by chance). The use of κ does carry some controversy largely due to the difficulty interpreting its level,[5,6] but it remains the most common metric.

Unfortunately, κ statistics and reliability coefficients cannot be calculated in studies without a negative case count. In this paper, we describe two alternative measures of interrater agreement that can work without a negative case count, and we show that they are in fact identical to each other.

Method

One way to quantify interrater agreement without a negative case count is to use the performance metrics commonly applied in information retrieval experiments.[7] There are two advantages to using these metrics: information retrieval researchers are familiar with them and they do not require a negative case count.

The two primary metrics are *precision* and *recall*. Given a subject and a gold standard, precision is the proportion of cases that the subject classified as positive that were positive in the gold standard. It is equivalent to positive predictive value. Recall is the proportion of positive cases in the gold standard that were classified as positive by the subject. It is equivalent to sensitivity. The two metrics are often combined as their

Affiliation of the authors: Department of Biomedical Informatics, Columbia University, New York, NY.

This work was funded by National Library of Medicine grants R01 LM06910 "Discovering and Applying Knowledge in Clinical Databases" and N01 LM07079 training grant.

Correspondence and reprints: George Hripcsak, MD, MS, Department of Medical Informatics, Columbia University, 622 West 168th Street, VC5, New York, NY 10032; e-mail: <hripcsak@columbia.edu>.

Received for publication: 11/03/04; accepted for publication: 01/04/05.

harmonic mean, known as the *F-measure*,[8] which can be formulated as follows:

$$F = \frac{(1+ß^2) \times \text{recall} \times \text{precision}}{(ß^2 \times \text{precision}) + \text{recall}} \quad (1)$$

ß allows one to weight either precision or recall more heavily, and they are balanced when ß = 1. In most experiments, there is no particular reason to favor precision or recall, so most researchers use ß = 1 (and Brants[7] used ß = 1 in the context of interrater agreement).

The agreement between two raters can be quantified using these traditional information retrieval metrics. Assume two raters perform the same task and agree as indicated in Table 1. Thus, a is the number of cases that both raters agree are positive, d is the number of cases that both raters agree are negative, and b and c are the number of cases that the raters disagree on.

Assume now that this is an information retrieval task and that d is unknown. In a phrase markup experiment, for example, it would represent the number of phrases that neither rater marked. The number of phrases marked by either rater (a, b, and c) can be counted, but because the total number of potential phrases in a document is poorly defined, the number of unmarked phrases is also undefined.

By treating one rater as the subject (say, rater 1) and temporarily treating the other rater's answers (rater 2's answers) as if they were a gold standard, one can calculate the precision as a/(a + b) and the recall as a/(a + c). The balanced F-measure is then given below:

$$F = \frac{2[a/(a+b)][a/(a+c)]}{a/(a+b)+a/(a+c)} \quad (2)$$

$$= 2a/(2a+b+c) \quad (3)$$

One can switch the roles so that rater 1's answers are temporarily treated as if they were a gold standard and rater 2 is treated as the subject. The resulting F-measure will be identical to Equation 3. Thus, it is does not matter which rater plays which role because the F-measure between them will be the same.[7]

The F-measure can be calculated in this way pairwise between all raters. The average F-measure among all pairs of raters can be used to quantify the agreement among the raters expressed as a familiar information retrieval measure. The higher the average F-measure, the more the raters agree with each other. If the raters are experts, and if the experts' answers are pooled to create a gold standard for some experiment, then the average F-measure among the experts quantifies their interrater agreement and therefore quantifies the gold standard's reliability. Unfortunately, the relation of this metric to traditional reliability theory is unclear.

Another approach is to apply traditional agreement measures. *Simple agreement* is defined as the proportion of cases for which a pair of raters agree, or (a + d)/(a + b + c + d).[9] Simple agreement requires knowing d. Even if d were known, it would not be useful with large d because it would approach one regardless of performance on positive cases (e.g., the agreement that most Internet documents are not relevant would overshadow the lack of agreement on relevant documents).

Table 1 ■ Agreement between Two Raters

		Rater 2's judgment	
		Positive	Negative
Rater 1's judgment	Positive	a	b
	Negative	c	d

Positive specific agreement[9] provides insight when the positive cases are rare.[6] It is the conditional probability that one rater will agree that a case is positive given that the other one rated it positive, where the role of the two raters is selected randomly. It approximates the proportion of positive cases that were agreed on. Positive specific agreement, p_{pos}, is defined as follows:

$$p_{pos} = 2a/(2a+b+c) \quad (4)$$

Equations 3 and 4 are identical. The balanced F-measure equals positive specific agreement. The overall agreement among the raters is simply the average positive specific agreement among all pairs of raters.

Investigators frequently use chance-corrected agreement, κ, to quantify reliability in classification experiments. κ is defined for Table 1 as follows[4]:

$$\kappa = \frac{2(ad-bc)}{(a+c)(c+d)+(b+d)(a+b)} \quad (5)$$

Calculating κ requires knowing d. If d is at least known to be large, however, the probability of chance agreement on positive cases approaches zero; Equation 5 approaches Equation 4, and κ approaches the positive specific agreement. Therefore, for experiments with large but unknown d, the average positive specific agreement, which equals the average F-measure among the raters, approaches the κ that would be calculated if d were known.

In text markup experiments, d (the number of phrases not marked) is poorly defined but not necessarily huge, and agreement can occur by chance. Average positive specific agreement (and therefore average F-measure among the raters) remains a good metric to report, although it will be higher than the chance corrected agreement (κ) that would be reported if d were knowable.

Discussion

In a typical information retrieval experiment, a computer or human subject performs some task and the subject's answers are compared with a gold standard generated by experts. The subject's performance is often reported as precision, recall, and F-measure, all of which can be calculated without a negative case count. The researcher is obligated to also report on the quality of the gold standard, yet the traditional interrater agreement measures like the κ statistic and others[5] cannot be calculated without the negative case count. This paper demonstrates that the average pairwise F-measure among the experts is equivalent to the average positive specific agreement among the experts, and it approaches the κ statistic that could be calculated if the negative case count were known.

An even better alternative is to carry out a separate measurement study[3] to assess the reliability of the gold standard. The

study should include a well-defined population of cases to assess the reliability of the expert.[6] The population should contain the kinds of cases that the investigator is hoping to differentiate (for example, depending on the goal, the investigator may want to differentiate nonrelevant from relevant documents or to differentiate somewhat relevant from very relevant). Such measurement studies can be constructed so that the cases are finite and well defined, the negative case count is known, and traditional reliability measures can be calculated. Resource constraints often hinder such measurement studies, unfortunately.

Some information retrieval studies include partial matches or other complications, and investigators frequently extend the definitions of precision and recall (e.g., half credit for a partial match). Quantifying the agreement among the gold standard raters becomes even more difficult, but an average F-measure using the extended precision and recall may suffice. A more advanced agreement model[6] is another alternative.

Our review of the literature revealed no previous reports on the correspondence among positive specific agreement, the F-measure, and κ. Graham and Bull[10] do report that positive specific agreement can be motivated as a weighted average of the sensitivities (recall) obtained when either rater is regarded as the gold standard. The latter is equivalent to the balanced F-measure.

In summary, in classification experiments in which the number of negative cases is unknown, undefined, or very large, one can quantify interrater agreement using the average positive specific agreement among the raters, which is identical to the average pairwise F-measure. This interrater agreement can be useful, for example, to quantify the reliability of an expert-derived gold standard. When the number of negative cases is large but unknown, these measures will approach the κ statistic, which is the traditional reliability metric that one would have calculated if the negative case count were known.

References ■

1. Hersh WR. Information retrieval: a health care prospective. New York: Springer; 1995, pp 45–50.
2. Hripcsak G, Wilcox A. Reference standards, judges, comparison subjects: roles for experts in evaluating system performance. J Am Med Inform Assoc. 2002;9:1–15.
3. Friedman CP, Wyatt JC. Evaluation methods in medical informatics. New York: Springer; 1997.
4. Fleiss JL. Statistical methods for rates and proportions. 2nd ed. New York: John Wiley & Sons; 1981; pp 212–36.
5. Uebersax JS. [cited 2005 March 23]. Available from: http://ourworld.compuserve.com/homepages/jsuebersax/agree.htm/.
6. Hripcsak G, Heitjan D. Measuring agreement in medical informatics reliability studies. J Biomed Inform. 2002;35:99–110.
7. Brants T. Inter-annotator agreement for a German newspaper corpus. Paper presented at the Second International Conference on Language Resources and Evaluation LREC-2000; 2000 May 31–June 2; Athens, Greece.
8. van Rijsbergen CJ. Information retrieval. 2nd ed. London: Butterworths; 1979.
9. Fleiss JL. Measuring agreement between two judges on the presence or absence of a trait. Biometrics. 1975;31:651–9.
10. Graham P, Bull B. Approximate standard errors and confidence intervals for indices of positive and negative agreement. J Clin Epidemiol. 1998;51:763–7.