

Performance and Knowledge of Performance

H5's Quality Assurance Protocol

Abstract

Large corporations are increasingly overwhelmed with electronic data, which undermines their ability to successfully meet their document preservation and production obligations. Document retrieval, classification or management tasks that could in the past be addressed through employee involvement or through manual attorney review are now prohibitively expensive or impossible to achieve. Increasingly, companies are turning to automated methods of information retrieval to meet their document preservation and production objectives. Central to the cost efficiency of an automated system is the system's ability to identify relevant records without simultaneously capturing a large amount of records that are off-target. Central to the defensibility of a company's selection and implementation of a solution is the measurement of the system's performance. A number of performance metrics for information retrieval, classification and review are occasionally proposed or espoused by industry participants: fallout, specificity, elusion, negative predictive value, accuracy (as a specific term of art), error rate, precision and recall. All metrics have some applicability in domains other than information retrieval. However, only precision and recall are adequate for purposes of information retrieval; other metrics would be unlikely to survive an expert challenge. This paper surveys the various metrics, explains why precision and recall are most meaningful in information retrieval, classification, and review performance, outlines H5's sample-based approach to their measurement, and explains why this approach enables H5 clients both to be *confident in the quality of H5's results* and to be *confident in their ability to demonstrate and defend the quality* of those results.

1. Introduction

Companies are finding that, in a world in which document populations grow at an ever-increasing rate and in which the array of laws and regulations that governs retention obligations grows increasingly complex, meeting document management and retrieval objectives is an increasingly challenging and risk-laden task. For assistance in meeting this challenge, companies are turning to a range of different approaches to document assessment, from variants of the traditional manual approach to highly automated systems. Any system that would provide a company with real assistance in carrying out the task not only must perform well (i.e., capture most of the target records without simultaneously capturing a large amount records that are off-target) but also must provide the company with evidence that it is performing well. Without valid and accurate knowledge of the performance of its chosen review system, a company will neither know for itself whether it has in fact met its retrieval obligations and objectives, nor will it be in a position to convince others (i.e., the court) that it has done so.

The purpose of this paper is to describe the protocol H5 follows in obtaining measures of the performance of its document retrieval system. H5 has conducted several studies, objectively monitored and validated, comparing the performance of its approach to document retrieval with that attained by a traditional approach. Each of these studies found that H5 significantly outperformed the manual approach with regard to both Recall and Precision, and H5 believes that these studies provide strong testimony of the performance capabilities of its process. H5 also believes, however, that those who would employ H5's assessment services (or any assessment product or service) have a right to understand the methodology by which the performance of those services would be measured on the specific project for which they would be contracted; only when equipped with this understanding will a company be able to assess whether H5's measurement protocol will suffice (a) to provide an accurate measure of retrieval performance and (b) to serve as a basis for defending the thoroughness and effectiveness of the company's retrieval efforts. This paper is intended to provide companies with that background.

The paper proceeds as follows. We begin with the question of the selection of performance metrics (Section 2); here, after reviewing the range of metrics available for measuring the performance of a document-assessment process, we identify the attributes of Recall and Precision that make them the metrics best suited to serve as measures of such a process. We then turn to the sample-based estimation of Recall and Precision (Section 3). The paper concludes with some observations on the implications of this measurement protocol, both for H5 and for those who engage H5's services (Section 4).

2. Selection of Metrics

A variety of metrics can be used to measure the performance of a document-assessment process. The first step in designing a protocol for measuring the effectiveness of a system is to decide which metrics will provide the most meaningful gauge of the system's performance. In this section, we begin with some background material that will lay the foundation for our further discussion of metrics (2.1); we then provide a brief review of the metrics most commonly used in the field of information retrieval (2.2); finally, we identify the attributes of Recall and Precision that we believe make them the most meaningful measures of the performance of a document-assessment process (2.3).

2.1 Background

We begin by establishing some points of reference for our discussion: a Venn diagram, a 2x2 table, and a concrete example.

A Venn diagram, like the following, provides a convenient way to visualize the results of a relevance review.

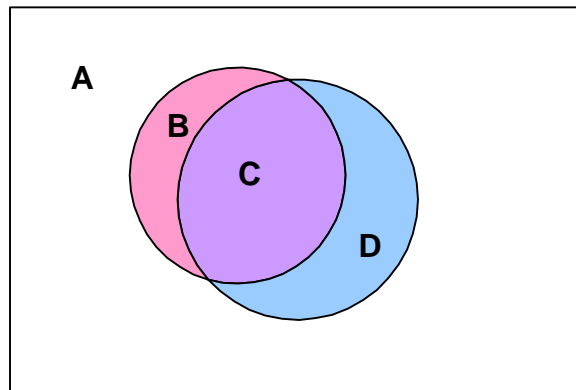


Figure 1: Venn Representation of Assessment Results

In the diagram above, we let the entire rectangle represent the full document population. The pink circle ($B \cup C$) represents the set of documents actually relevant to the target category; the blue circle ($C \cup D$) represents the set of documents the process under inspection has assessed as relevant to the category; the intersection of the two circles (the purple lens) represents the subset of relevant documents that the process has succeeded in identifying as relevant. More specifically, the elements of the diagram are defined as follows:

- $A \cup B \cup C \cup D$: the full document population;
- $B \cup C$: the set of documents actually relevant to the target category;
- $C \cup D$: the set of documents a process has assessed as relevant to the target category;
- **A**: the set of documents that are not relevant and that have been assessed as not relevant;
- **B**: the set of documents that are relevant but that the process has failed to assess as relevant;
- **C**: the set of documents that are relevant and that the process has succeeded in assessing as relevant;

- **D**: the set of documents that are not relevant but that the process has (incorrectly) assessed as relevant.

In a relevance review, the goal is, generally, to obtain a result in which the set of documents assessed as relevant (the blue circle) overlaps as closely as possible with the set of documents that are actually relevant (the pink circle); put another way, the objective is to maximize Set C while minimizing Sets B and D.

The results of a relevance review can also be conveniently summarized as a 2x2 table, a format that proves to be useful for the derivation of performance metrics.

Table 1: 2x2 Table of Assessment Results

		Actual Status		Total
		Relevant	Not Rel	
Assessed Status	Relevant	n_{11}	n_{12}	n_{1+}
	Not Rel	n_{21}	n_{22}	n_{2+}
Total		n_{+1}	n_{+2}	n_{++}

In the table, the count n_{11} represents the number of documents that were assessed as relevant and that were actually relevant (i.e., the count of documents contained in Set C of the Venn diagram); the count n_{12} represents the number of documents that were assessed as relevant but were not actually relevant (i.e., the count of documents in Set D); the count n_{21} represents the number of documents that were assessed as not relevant but were actually relevant (i.e., the count of documents in Set B); and the count n_{22} represents the number of documents that were assessed as not relevant and that were actually not relevant (i.e., the count of documents in Set A). The remaining values in the table represent the marginal sums for the rows (n_{1+} , n_{2+} ; i.e., total documents assessed as relevant and total documents assessed as not relevant), for the columns (n_{+1} , n_{+2} ; i.e., total documents actually relevant and total documents actually not relevant), and for the table as a whole (n_{++} ; i.e., total documents in the population).

In a relevance review, the goal is to maximize the counts in the upper-left and lower-right cells of the 2x2 table (n_{11} and n_{22}) and to minimize the counts in the lower-left and upper-right cells (n_{21} and n_{12}).

To take a hypothetical example, suppose that a population of 1,000,000 documents were reviewed for relevance to a particular category and suppose that, in that population, 10,000 documents (1% of the population) were actually relevant to the category; suppose further that a relevance review of the population assessed 12,000 documents as relevant, 8,000 of which were actually relevant and 4,000 of which were actually not relevant. The results of the review could then be summarized in the following table.

Table 2: Example of 2x2 Summary of Assessment Results

		Actual Status		Total
		Relevant	Not Rel	
Assessed Status	Relevant	8,000	4,000	12,000
	Not Rel	2,000	986,000	988,000
Total		10,000	990,000	1,000,000

In terms of the Venn diagram, the hypothetical example means that Set A (the white area) contains 986,000 documents, Set B (the pink crescent) contains 2,000 documents, Set C (the purple lens) contains 8,000 documents, and Set D (the blue crescent) contains 4,000 documents.

With these points of reference established, a review of standard IR metrics, their definitions and the aspects of performance they quantify, is a straightforward matter.

2.2 Inventory of Metrics

Most metrics used to quantify the performance of IR systems are proportions (or percentages) of marginal totals; in addition to these, a few metrics are designed to quantify the overall performance of a system. We begin by reviewing the first sort, the proportions, and then turn to the composite measures.

One (complementary) pair of metrics condition on the total number of actually relevant documents (n_{+1}).

- **Recall (= Sensitivity)**

- $n_{11} / n_{+1} [= C / (C + B)]$ (1)
- E.g., $8000 / (8000 + 2000) = 0.800 (= 80.0\%)$
- Recall (or “Sensitivity”, as the measure is called in the field of diagnostic test evaluation) is a measure of how completely a process has captured the target set of relevant documents. For a company, achieving high Recall on a document-assessment project is crucial because the failure to do so may mean a failure to meet its document-retrieval objectives and obligations (and a failure to be able to argue, defensibly, that it has met those obligations).

- **False Negative Rate**

- $n_{21} / n_{+1} [= B / (C + B)]$ (2)
- E.g., $2000 / (8000 + 2000) = 0.200 (= 20.0\%)$
- False Negative Rate, a metric used more often in the field of diagnostic test evaluation than in IR, measures how much of the target set has been missed by a process. The complement to Recall, it measures the same aspect of performance as does that metric, but from the opposite perspective (i.e., how incompletely the target set has been captured).

Another pair of metrics condition on the total number of actually non-relevant documents (n_{+2}).

- **Fallout (= False Positive Rate)**

- $n_{12} / n_{+2} [= D / (D + A)]$ (3)
- E.g., $4000 / (4000 + 986000) = 0.004 (= 0.4\%)$
- Fallout (or “False Positive Rate”, as the measure is known in the field of diagnostic test evaluation) is a measure of how much of the actually non-relevant set has been incorrectly assessed as relevant. In the IR domain, Fallout is largely a secondary metric, providing supplementary context for evaluating the level of Precision (see below) achieved by a system. In a retention project, for example, Fallout could provide a useful gauge of the material impact a low level of Precision could have on the ability to purge non-relevant material.

- **Specificity**

- $n_{22} / n_{+2} [= A / (D + A)]$ (4)
- E.g., $986000 / (4000 + 986000) = 0.996 (= 99.6\%)$
- Specificity, another metric used more often in the field of diagnostic test evaluation than in IR, measures how much of the actually non-relevant set has been correctly assessed as not relevant. Like its complement, Fallout, Specificity is, in IR, a secondary metric, supplying supplemental information when the focus of a project is on expiring unwanted material.

Another pair condition on the total number of documents assessed as relevant (n_{1+}).

- **Precision (= Positive Predictive Value)**

- $n_{11} / n_{1+} [= C / (C + D)]$ (5)
- E.g., $8000 / (8000 + 4000) = 0.667 (= 66.7\%)$
- Precision (or “Positive Predictive Value,” as the measure is called in the field of diagnostic test evaluation) is a measure of how on-target a system’s assessments are. For a company, achieving high Precision on a document-review project is crucial because the failure to do so may mean both a wasteful and ineffective allocation of time and resources and the retention of potentially troublesome records that could otherwise safely and legally be purged.

- **Imprecision**

- $n_{12} / n_{1+} [= D / (C + D)]$ (6)
- E.g., $4000 / (8000 + 4000) = 0.333 (= 33.3\%)$
- Imprecision, as we might call the complement to Precision, measures the same aspect of performance as does Precision, just from the opposite perspective (i.e., how off-target a system’s relevance assessments are).

A fourth pair condition on the total number of documents assessed as not relevant (n_{2+}).

- **Elusion**

- $n_{21} / n_{2+} [= B / (B + A)]$ (7)
- E.g., $2000 / (2000 + 986000) = 0.002 (= 0.2\%)$
- Elusion, the complement to Negative Predictive Value, is a measure of how many actually relevant documents remain in the set a system has assessed as not relevant. While potentially useful in some applications (e.g., in an acceptance testing scenario), the metric has significance only when used in conjunction with other metrics (more specifically, the rate of false negatives in the part of the population assessed as not relevant admits of meaningful interpretation only when we know the proportion of actually relevant documents in the full population).¹

- **Negative Predictive Value**

- $n_{22} / n_{2+} [= A / (B + A)]$ (8)
- E.g., $986000 / (2000 + 986000) = 0.998 (= 99.8\%)$
- Negative Predictive Value, a metric used more often in the field of diagnostic test evaluation than in IR, measures how on-target a system’s negative (not relevant) assessments are. Like other proportional metrics that condition on either the total number of documents assessed as not relevant (Elusion) or the total number of documents that are actually non-relevant (Fallout and Specificity), the range of values this metric is likely to take is heavily influenced by the proportion of relevant to non-relevant material: when non-relevant material represents a high proportion of a population (as is almost always the case in a litigation or retention scenario), Negative Predictive Value will almost always take on high values.

Finally, some metrics condition on the total number of documents in the population (n_{++}).

- **Accuracy**

- $(n_{11} + n_{22}) / n_{++} [= (C + A) / (A + B + C + D)]$ (9)
- E.g., $(8000 + 986000) / (986000 + 2000 + 8000 + 4000) = 0.994 (= 99.4\%)$

¹ And putting the rate in that context brings us back, in the end, to Recall.

- Accuracy is a measure of how correct, in aggregate, a system's assessments are. In practice, the Accuracy metric is almost never meaningful, and is often misleading, as it is dominated by the usually very large subset of non-relevant documents that a system has (correctly) not assessed as relevant. When the yield of relevant documents is low, a system can get a very high Accuracy score simply by assessing nothing as relevant (that is, a system can achieve high Accuracy even while achieving 0% Recall). For this reason, Accuracy, in the precise sense defined here, is rarely used as a measure of the performance of IR systems.
- **Error**
 - $(n_{12} + n_{21}) / n_{++} [= (D + B) / (A + B + C + D)]$ (10)
 - E.g., $(4000 + 2000) / (986000 + 2000 + 8000 + 4000) = 0.006 (= 0.6\%)$
 - Error is a measure of how incorrect, in aggregate, a system's assessments are. Like its complement, Accuracy, and for the same reasons, Error is almost never meaningful and is often misleading. Error is rarely used as a measure of the performance of IR systems.

In addition to the above metrics, which quantify specific aspects of the performance of a system, some metrics are designed to gauge the overall performance of a system. Three such metrics are the following

- **F-Measure**
 - $(2 \times \text{Recall} \times \text{Precision}) / (\text{Recall} + \text{Precision})$ (11)
 - E.g., $(2 \times 0.8 \times 0.667) / (0.8 + 0.667) = 0.727 (= 72.7\%)$
 - The F-Measure, a function of the aspects of performance that are measured separately by Recall and Precision, is intended to provide a single measure of the overall performance of a system. When Recall and Precision are given equal weight, the F-Measure resolves to the harmonic mean of the two metrics (as in the formula above); it is also possible, however, to make use of variants of the F-Measure that give unequal weights to Recall and Precision, should project requirements dictate such differences in performance priorities. While helpful as an overall measure, and so perhaps useful in comparative studies, the F-Measure, by its composite nature, obscures the view of the specific aspects of performance that the more narrowly targeted metrics Recall and Precision usefully bring into view.
- **Odds Ratio**
 - $(n_{11} / n_{12}) / (n_{21} / n_{22}) = (n_{11}n_{22}) / (n_{21}n_{12}) = (C \times A) / (B \times D)$ (12)
 - E.g., $(8000 / 4000) / (2000 / 986000) = 986.0$
 - The odds ratio, another measure of overall performance, quantifies the difference between the odds of finding a relevant document in the set assessed as relevant and the odds of hitting a relevant document in the set assessed as not relevant. When a system performs well (i.e., succeeds in segregating the relevant from the non-relevant), that difference in odds will be great (as is the case in the running example we have been using); when a system performs poorly (i.e., fails to segregate the relevant from the non-relevant) the difference in odds will be small (i.e., the odds ratio will approach 1.0). While the odds ratio is attractive from a number of standpoints (especially for comparative purposes and for more complex modeling of performance), it, like the other aggregate metrics, has the drawback that it obscures our view of certain specific aspects of performance that are often of interest in themselves.

The above list is by no means an exhaustive inventory of the metrics used in assessing the performance of information retrieval systems (and processes designed for like tasks). There are, for example, variants of some of the above that can be applied when a system produces a ranked set of results (e.g., Average Precision, which is sometimes used in TREC studies).² The above list does, however, provide a

2 H5's document-assessment process is designed to produce binary assessments of relevance to categories that have been precisely defined by the company or law firm that has engaged H5's services. While H5 does offer a

comprehensive view of the best candidates for measuring the performance of a document-assessment process, like H5's, in its usual application.

2.3 Selection of Metrics

In looking at the results of a relevance-assessment effort (whether it be for retention, production, or issue-coding purposes), the attorney who initiated the effort will want to know the answer to two primary questions: (1) Does the result set capture a high proportion of the material I am looking for (and so put me in a position to meet my obligations and objectives)? (2) Does the result set minimize the amount of non-relevant material I have to review or retain (thereby saving me from a wasteful and ineffective review and, in a retention scenario, allowing me to make a more complete purge of unwanted material)?

Of the metrics reviewed above, Recall and Precision provide the most direct and meaningful answers to these questions. In the hypothetical scenario we have been using as a running example, Recall tells us that we have captured 80% of the material we set out to find; Precision tells us that 66.7% of the material we have assessed to be relevant is actually relevant. Because Recall and Precision condition on the marginal sums directly pertinent to the primary questions identified above (total actually relevant documents, total documents assessed as relevant), they serve as the most sensitive gauges of the aspects of performance that are the focus of the questions.

The metrics that condition on the non-relevant marginal totals (whether total actually non-relevant documents or total documents assessed as non-relevant), are not sensitive to the aspects of performance of primary interest to those who initiate a review. The values taken by these metrics, moreover, are strongly influenced by the disproportionately large number of non-relevant documents in a typical review population (most of which any process will leave in the not relevant set). In our running example, the hypothesized process achieves a Specificity of 99.6% and a Negative Predictive Value of 99.8%; unless used in conjunction with other measures (in particular, in conjunction with the yield of relevant documents), these values will fail to bring to light that the process missed one relevant document for every four that it captured and returned one non-relevant document for every two relevant documents.

The Accuracy metric (like the complementary Error metric) is not a meaningful gauge of performance, both because it is strongly influenced by the usually disproportionately large number of non-relevant documents (note, in the running example, the achievement of 99.4% Accuracy even when Recall was at 80% and Precision below 70%) and because it does not provide a focused view into key aspects of performance: how many relevant documents have been missed (Type II error, which is gauged by Recall) and how many non-relevant documents have been given a positive assessment (Type I error, which is gauged by Precision).

Of the overall metrics, the mathematical properties of the Odds Ratio make it the most attractive, especially when comparing the performance of different systems. Its unitary nature, however, though a strength when comparing systems, can be a weakness when assessing performance in practice. When we are assessing the performance of a system and looking for ways we can improve on performance, we generally want separate views into how much Type I error and how much Type II error the system is generating; the Odds Ratio does not give us those views.

H5 believes that Recall and Precision provide the most meaningful measures of the aspects of performance that matter most in a relevance-assessment project, and these are the metrics that H5 places at the heart of its quality control processes. Other metrics can provide useful alternative views of performance and should be used when circumstances and project objectives warrant, but the perspectives these metrics provide are supplemental to those provided by Recall and Precision. In the remainder of this paper, we will discuss the protocol H5 follows in obtaining valid measures of Recall and Precision.

number of options for prioritizing results, its document-assessment process does not produce a continuously ranked set of results.

3. Sample-Based Measurement of Recall and Precision

Looking back at the formulae given above for Recall and Precision, (1) and (5) respectively, we see that the calculation of Recall and Precision is quite straightforward, as long as we know the input values (the numbers of documents in Sets B, C, and D of Figure 1 or the counts n_{11} , n_{12} , and n_{21} in Table 1). The calculation of the metrics becomes challenging, however, when we do not, and cannot, know the true input values (the true sizes of the sets), and that is almost always the condition in which we find ourselves in the real world.

In a real-world document-review project, the size of the document population under review will almost always be so large as to preclude knowledge of the exact counts of documents in each of the subsets pertinent to the calculation of Recall and Precision. The document populations that H5 is asked to assess for relevance typically number in the millions or tens of millions of documents and, at current rates of growth in document populations, it is safe to predict that these numbers will seem small in the near future. In these circumstances, it is not feasible to obtain knowledge of the exact values of the inputs to Recall and Precision (and therefore not possible to obtain knowledge of the exact values of the metrics themselves). It is possible, however, to obtain statistically-valid estimates of the values of those inputs (and thereby to obtain statistically-valid estimates of the metrics). In this section, we review the procedures by which H5 uses sampling to obtain statistically-valid estimates of Recall and Precision.

There is more than one way to obtain estimates of Recall and Precision. In this section, we consider, from a high level, the approach H5 follows in obtaining estimates of the metrics. We begin by reviewing one very intuitive way of going about the task and then identifying the drawbacks that make that approach less attractive as a way of measuring performance on an actual document retrieval project (3.1). We then provide a conceptual overview of the approach H5 does take to meeting its data collection and performance measurement objectives (3.2).

3.1 Test Samples

Test samples, samples selected and reviewed solely for the purpose of measuring performance, offer an intuitive and straightforward way of obtaining estimates of Recall and Precision. The test-sample approach can be implemented in a few different ways; a sketch of one implementation follows.

- 1) Once development of the assessment software has reached a certain level of maturity (i.e., a level at which the results of testing will be meaningful), a simple random sample of the full document population is selected; the size of the sample selected will be determined by the expected yield of relevant documents and by the amount of uncertainty one is willing to tolerate in one's estimates.
- 2) The sample is then reviewed manually, with documents being coded as relevant or not relevant. Because these assessments will be the standard against which the performance of the software will be measured, it will be necessary to implement appropriate quality control procedures in order to ensure that the manual assessments are indeed accurate.
- 3) The assessment software to be tested is applied to the sample, and the results are cross-classified with those of the manual review of the sample. The result is a 2x2 table like the following.

Table 2: Test-Sample Results

		Manual Review		Total
		Relevant	Not Rel	
Software Review	Relevant	n_{11}	n_{12}	n_{1+}
	Not Rel	n_{21}	n_{22}	n_{2+}
Total		n_{+1}	n_{+2}	n_{++}

- 4) Analysis of the table follows in a straightforward manner, with estimates and confidence intervals for Recall (R_m) and Precision (P_m) being derived, assuming a multinomial sampling design, as follows.

$$\text{Recall Estimate: } \hat{R}_m = n_{11} / n_{+1} \quad (14)$$

$$\text{Estimated Variance: } \hat{\text{var}}(\hat{R}_m) = \frac{\hat{R}_m(1 - \hat{R}_m)}{n_{+1} - 1} \quad (15)$$

$$\text{90\% Confidence Interval: } \hat{R}_m \pm 1.645\sqrt{\hat{\text{var}}(\hat{R}_m)} \quad (16)$$

$$\text{Precision Estimate: } \hat{P}_m = n_{11} / n_{1+} \quad (17)$$

$$\text{Estimated Variance: } \hat{\text{var}}(\hat{P}_m) = \frac{\hat{P}_m(1 - \hat{P}_m)}{n_{1+} - 1} \quad (18)$$

$$\text{90\% Confidence Interval: } \hat{P}_m \pm 1.645\sqrt{\hat{\text{var}}(\hat{P}_m)} \quad (19)$$

While straightforward, and perhaps useful in some circumstances, the test-sample approach does have significant limitations. Chief of these are the following:

- In order to obtain precise estimates of Recall and Precision, relatively large samples will have to be drawn and reviewed, with the required sample size increasing as the yield of relevant material decreases. For example, for a category with an 8% yield, in order to obtain, for Recall, a 90% confidence interval with a half-width of 5%, one would have to select and review a sample of 2,200 documents;³ for a category with a 3% yield, achieving the same goal would require a sample of 5,800 documents; for a category with a 1% yield, the size of sample required to meet those objectives would be 17,200 documents.
- The test sample serves performance measurement purposes only; it does not serve (or serves only in a limited capacity) data-collection purposes. In order to develop assessment software that succeeds in performing well, additional samples will have to be drawn and reviewed in order to gather the input data needed for effective development.
- Even if used, post-test, for data collection purposes, the yield of useful data will in most cases be small, as the sampling design does not allow disproportionate targeting of relevant material or the targeting of specific parts of the document collection.
- Each time the development team wants to gauge performance, a new sample will have to be drawn and reviewed, limiting the amount of testing that can be done and making real-time monitoring of performance prohibitively expensive.
- The limitations above, which pertain primarily to efficiency with which the desired information is acquired, are compounded if one wishes to measure performance, not at the aggregate level, but at a category-specific level.

A sampling protocol that will meet the needs of a team developing relevance-assessment software must satisfy a number of requirements. (1) It must enable the team to collect the data it needs to develop software that performs well. (2) It must enable the team to collect the data it needs to know that the software is performing well (or poorly). (3) It must enable the team to monitor the performance of the

³ Assuming, for the purpose of this example, that the actual levels of Recall and Precision achieved were both 80%.

software on a real-time basis, allowing development decisions to be based on then-current levels of performance. (4) It must provide the team with the preceding capabilities without requiring an undue expenditure of time or resources.

H5 believes that the test-sample approach, though not without application in certain circumstances, is unable, due to the limitations noted above, to meet, at any tolerable level of expenditure of time and resources, the information and data requirements of a team developing automated assessment software. For this reason, H5 has designed, and makes use of, an alternative approach, a yield-based approach, for collecting input data and obtaining performance metrics. We now turn to a description of that approach, beginning with a conceptual overview.

3.2 Yield-Based Estimation

In contrast to the test-sample approach just described, H5 takes a yield-based approach to measuring performance. In this section, we provide a conceptual overview of the approach. The primary focus of the overview will be on the procedures by which we obtain estimates of Recall (the harder of the two metrics to estimate); as will be seen, we obtain estimates of Precision (the easier of the two metrics to estimate) in the course of calculating Recall.

We begin by returning to formula (1), the basic formula for Recall, repeated here for convenience.

$$\text{Recall} = n_{11} / n_{+1} [= C / (C + B)] \quad (1)$$

From the formula, we see that Recall is a function of two values: (a) the number of actually relevant documents that have been correctly assessed as relevant (n_{11} , or the count of documents in Set C) and (b) the total number of actually relevant documents (n_{+1} , or the count of documents in the set formed by the union of Sets C and B). Neither of these values, in any but the smallest of modern-day document-assessment projects, is likely to be known or knowable. Each of the values can, however, be estimated through the selection and review of samples, and, by combining these component estimates in the appropriate way, we can obtain a valid estimate of the metric that is a function of them, Recall. We now step through, at a high level, the procedures H5 follows in obtaining an estimate of n_{+1} (the yield estimate), in obtaining an estimate of n_{11} (the estimate of the number of correct positive assessments), and in combining these two estimates to obtain an estimate of Recall.

3.2.1 Yield Estimate

The procedures H5 follows in obtaining an estimate of the yield are, in concept, quite straightforward: (1) we take a valid sample from the full target population; (2) we conduct a quality-controlled manual review of the sample, coding sample documents as relevant or not relevant; (3) we project from observed counts in the sample to estimated yield in the full population. Taking this approach, if a Simple Random Sampling design were used, the yield estimate would be calculated as follows.

For the following, let:

N = total documents in the full population (i.e., = n_{++});

τ = total relevant documents in the full population (i.e., = n_{+1});

p = proportion of full population that is relevant;

n = total documents in the sample; and

a = total relevant documents observed in the sample.

Our estimate of total relevant documents is then calculated as follows.

$$\text{Estimated proportion: } \hat{p} = \frac{a}{n} \tag{20}$$

$$\text{Estimated total: } \hat{\tau} = N\hat{p} \tag{21}$$

$$\text{Sample Variance: } s^2 = \left(\frac{n}{n-1}\right)\hat{p}(1-\hat{p}) \tag{22}$$

$$\text{Estimated Variance of Estimator: } \hat{\text{var}}(\hat{\tau}) = N(N-n)\left(\frac{s^2}{n}\right) \tag{23}$$

$$\text{90\% Confidence Interval: } \hat{\tau} \pm 1.645\sqrt{\hat{\text{var}}(\hat{\tau})} \tag{24}$$

Now, in fact, H5 does not use a Simple Random Sampling design to collect the data on which the estimate of yield is based; H5, for efficiency reasons, uses an adaptive form of disproportionate stratified sampling for this purpose. The particular sampling design, however, is not of concern at the moment; what is important to note here is that one component of our estimate of recall is a sample-based estimate of the yield of relevant documents.

Put in terms of a 2x2 table, this step of the procedure can be thought of as populating the table's column marginal sums with sample-based estimates. We begin with a known value for the population total (n_{++}).

Table 3: Yield Estimate – Initial State

		Manual Review		Total
		Relevant	Not Rel	
Software Review	Relevant			
	Not Rel			
Total				n_{++}

We then use the results of a review of a sample of the full population to obtain estimates of column marginal proportions ($\hat{p}_{+1}, \hat{p}_{+2}$).

Table 4: Yield Estimate – Estimate of Marginal Proportions

		Manual Review		Total
		Relevant	Not Rel	
Software Review	Relevant			
	Not Rel			
Total		\hat{p}_{+1}	\hat{p}_{+2}	1.0

From these estimates we obtain estimates of the marginal sums ($\hat{n}_{+1}, \hat{n}_{+2}$).

Table 5: Yield Estimate – Estimate of Marginal Sums

		Manual Review		Total
		Relevant	Not Rel	
Software Review	Relevant			
	Not Rel			
Total		\hat{n}_{+1}	\hat{n}_{+2}	n_{++}

The estimate of the marginal sum \hat{n}_{+1} is equal to $\hat{\tau}$ as used in the formulae above.

3.2.2 Estimate of Correct Positive Assessments

The procedures H5 follows in obtaining an estimate of the number of correct positive assessments are also quite straightforward. In order to obtain this estimate, (1) we apply the software in its then-current state to the full population, tagging population documents as relevant or not relevant; (2) we take a valid sample from the set of documents that have been tagged as relevant; (3) we conduct a quality-controlled manual review of the sample, coding sample documents as actually relevant or actually non-relevant; (4) we project from observed correct positive software assessments in the sample to an estimated number of correct positive assessments in the full population. Using a Simple Random Sampling design, the yield estimate would be calculated as follows.

For the following, let:

- N_p = total positive software assessments in the full population (i.e., = n_{1+});
- τ_c = total correct positive assessments in the full population (i.e., = n_{11});
- p_c = proportion of positive assessments in the full population that are correct;
- n_p = total documents in the sample of positive assessments; and
- a_c = total correct positive assessments observed in the sample.

Our estimate of total correct positive assessments is then calculated as follows.

$$\text{Estimated proportion: } \hat{p}_c = \frac{a_c}{n_p} \tag{25}$$

$$\text{Estimated total: } \hat{\tau}_c = N_p \hat{p}_c \tag{26}$$

$$\text{Sample Variance: } s^2 = \left(\frac{n_p}{n_p - 1} \right) \hat{p}_c (1 - \hat{p}_c) \tag{27}$$

$$\text{Estimated Variance of Estimator: } \hat{\text{var}}(\hat{\tau}_c) = N_p (N_p - n_p) \left(\frac{s^2}{n_p} \right) \tag{28}$$

$$\text{90\% Confidence Interval: } \hat{\tau}_c \pm 1.645 \sqrt{\hat{\text{var}}(\hat{\tau}_c)} \tag{29}$$

As was the case with the yield estimate, conditions may sometimes dictate the use of a sampling design other than that assumed here; for example, if the full population is very large, we may, for efficiency reasons, choose to apply the software to a (large) sample of the full population and then draw a review sample from the positive assessments made on that initial sample. For purposes of this overview, however, what is important to note is simply that the second component of our estimate of recall is a sample-based estimate of the number of correct positive assessments.

Put in terms of a 2x2 table, this step of the procedure can be thought of as populating the table's row 1 counts with sample-based estimates. We begin with a known value for the row 1 marginal sum (n_{1+}).

Table 6: Correct Positive Assessment Estimate – Initial State

		Manual Review		Total
		Relevant	Not Rel	
Software Review	Relevant			n_{1+}
	Not Rel			
Total				

We then use the results of a review of a sample of row 1 documents to obtain estimates of row 1 conditional proportions ($\hat{p}_{1|1}$, $\hat{p}_{2|1}$).

Table 7: Estimate of Conditional Proportions

		Manual Review		Total
		Relevant	Not Rel	
Software Review	Relevant	$\hat{p}_{1 1}$	$\hat{p}_{2 1}$	p_{1+}
	Not Rel			
Total				

From these estimates we obtain estimates of the row 1 counts (\hat{n}_{11} , \hat{n}_{12}).

Table 8: Estimate of Cell Counts

		Manual Review		Total
		Relevant	Not Rel	
Software Review	Relevant	\hat{n}_{11}	\hat{n}_{12}	n_{1+}
	Not Rel			
Total				

The estimate of the cell count \hat{n}_{11} is equal to $\hat{\tau}_c$ as used in the formulae above.

3.2.3 Estimates of Recall and Precision

With estimates of yield and of total correct positive assessments in hand, the calculation of a yield-based estimate for Recall (R_y) follows in a straightforward manner.

$$\text{Recall Estimate: } \hat{R}_y = \frac{\hat{\tau}_c}{\hat{\tau}} \tag{30}$$

$$\text{Estimated Variance of Estimator:}^4 \text{ } \hat{\text{var}}(\hat{R}_y) = \hat{R}_y^2 \left(\frac{\hat{\text{var}}(\hat{\tau}_c)}{\hat{\tau}_c^2} + \frac{\hat{\text{var}}(\hat{\tau})}{\hat{\tau}^2} \right) \tag{31}$$

$$\text{90\% Confidence Interval: } \hat{R}_y \pm 1.645 \sqrt{\hat{\text{var}}(\hat{R}_y)} \tag{32}$$

The components of an estimate of Precision (P_y) were gathered in the course of obtaining our estimate of the total number of correct positive assessments; the metric is calculated as follows.

$$\text{Precision Estimate: } \hat{P}_y = \hat{p}_c = \frac{a_c}{n_p} \tag{33}$$

4 The variance for the Recall estimate is calculated by applying the principles of Gaussian error propagation. In formula (31), we set the covariance between n_{11} and n_{+1} to zero, on the grounds that the estimates for those values are obtained through independent sampling and therefore the variances associated with those estimates (what matters for error propagation) are independent.

$$\text{Estimated Variance of Estimator: } \hat{\text{var}}(\hat{P}_y) = \left(\frac{N_p - n_p}{N_p} \right) \frac{\hat{p}_c(1 - \hat{p}_c)}{n_p - 1} \quad (34)$$

$$90\% \text{ Confidence Interval: } \hat{P}_y \pm 1.645 \sqrt{\hat{\text{var}}(\hat{P}_y)} \quad (35)$$

Put in terms of a 2x2 table, this step in the procedure is just a matter of using already-derived estimates to obtain estimates of the pertinent conditional proportions. For Recall, that means finding the conditional proportion $\hat{n}_{11} / \hat{n}_{+1}$.⁵

Table 9: Estimate of Recall

		Manual Review		Total
		Relevant	Not Rel	
Software Review	Relevant	\hat{n}_{11}	\hat{n}_{12}	n_{1+}
	Not Rel	\hat{n}_{21}	\hat{n}_{22}	n_{2+}
Total		\hat{n}_{+1}	\hat{n}_{+2}	n_{++}

For Precision, that means finding the conditional proportion \hat{n}_{11} / n_{1+} .

Table 9: Estimate of Precision

		Manual Review		Total
		Relevant	Not Rel	
Software Review	Relevant	\hat{n}_{11}	\hat{n}_{12}	n_{1+}
	Not Rel	\hat{n}_{21}	\hat{n}_{22}	n_{2+}
Total		\hat{n}_{+1}	\hat{n}_{+2}	n_{++}

3.2.4 Example

As an illustration, suppose we had a full population ($= N$) of 1,000,000 documents, and suppose that we reviewed a simple random sample ($= n$) of 5,000 documents from the population and, in doing so, observed a total ($= a$) of 500 relevant documents in the sample. From these data, we would estimate that 10% of the full population were relevant; in terms of documents, this would come to an estimate of 100,000 relevant documents, with a 90% confidence interval of (93,038, 106,962).

Suppose then that, at some stage of development, we applied the software, in its then-current state, to the full population and tagged a total ($= N_p$) of 94,118 documents as relevant; suppose as well that we reviewed a simple random sample ($= n_p$) of 300 documents from the full set of tagged documents and, in doing so, observed a total ($= a_c$) of 255 actually relevant documents in the sample. From these data we would estimate that we had correctly tagged as relevant a total of 80,000 documents, with a 90% confidence interval of (76,808, 83,192).

We would then be in a position, by putting together the results of our two sampling tracks, to obtain estimates of Recall and Precision. Our estimate of Recall is 80.0%, with a 90% confidence interval of (73.6%, 86.4%). Our estimate of Precision is 85.0%, with a 90% confidence interval of (81.6%, 88.4%).

⁵ Note that, from the estimates already obtained, we are able to populate the remainder of the table.

3.2.5 Summary

Viewed from a high level, H5’s yield-based approach to measuring Recall is reasonably intuitive: we follow separate sampling tracks in order to obtain estimates of each of the components of Recall and then synthesize the results in order to obtain an estimate of the metric itself. An estimate of Precision is obtained as a matter of course in following the protocol for estimating Recall. Figure 2 below summarizes, in terms of a Venn diagram, the procedures H5 follows.

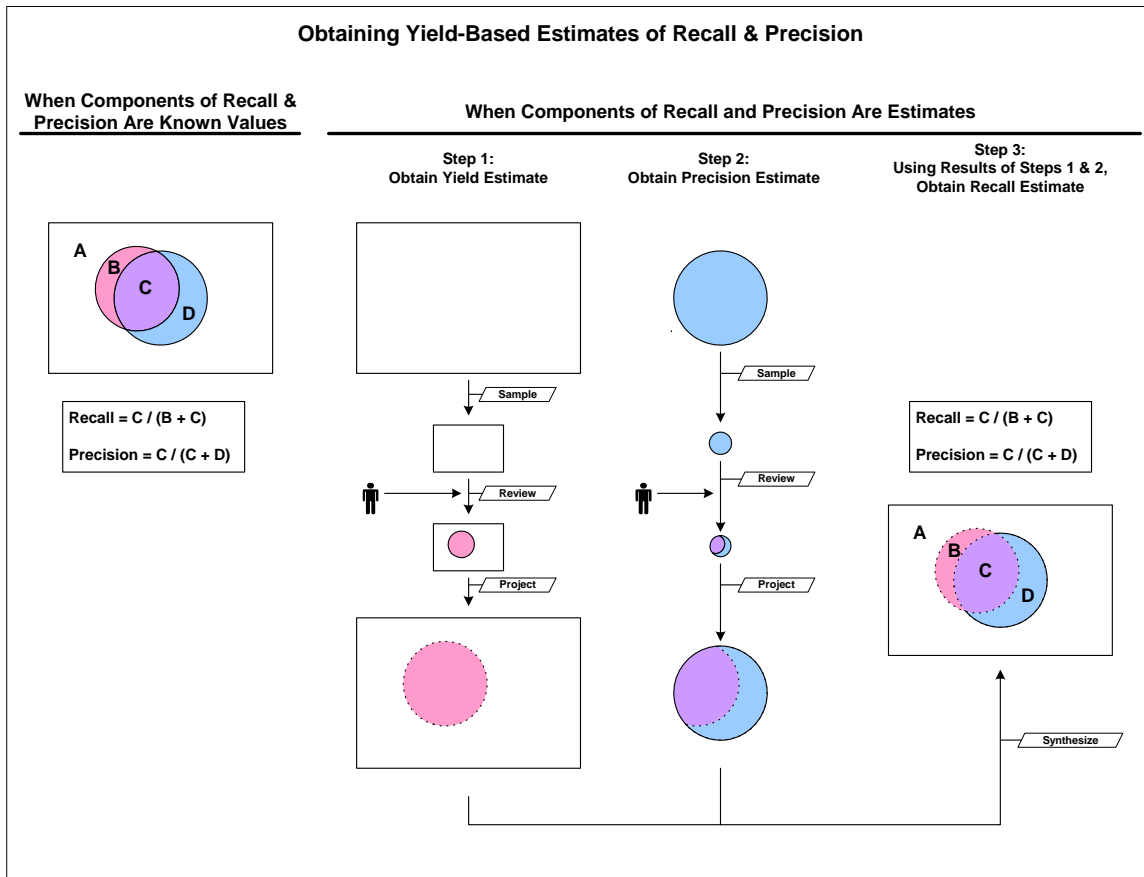


Figure 2: Yield-Based Approach to the Measurement of Recall

4. Conclusion – Implications for H5 and H5’s Clients

The yield-based measurement protocol we have just reviewed has important implications both for H5 and for H5’s clients. In this section, we consider how the protocol enables H5 to ensure the quality of the output of its document-assessment process and how the protocol enables H5’s clients to be confident of the quality and defensibility of H5’s results.

4.1 Implications for H5’s QA processes

The measurement protocol just reviewed provides H5 with three capabilities that are essential to its quality assurance regimen: (1) valid and precise estimates of recall and precision, (2) aggregate and topic-specific metrics, and (3) real-time measures of performance over the course of a project. A few notes on the implications of these capabilities for H5’s quality assurance program follow.

- **Valid and precise estimates of Recall and Precision.** We noted earlier that the metrics that are the most sensitive and informative gauges of the effectiveness of a document-retrieval effort are Recall and Precision: Recall, because it quantifies the extent to which a process has found all that you either need or want to find; Precision, because it quantifies the extent to which a process has avoided the inclusion of unwanted or non-relevant material in your result set. H5's measurement protocol, by providing an H5 project team with precise and statistically-valid estimates of Recall and Precision, enables H5 to focus its quality control processes on the aspects of performance that matter most in a document-assessment effort.
- **Aggregate and topic-specific metrics.** The protocol just reviewed is suitable for obtaining both aggregate measures of performance (i.e., the Recall and Precision achieved across all topics or categories) and topic-specific measures of performance (i.e., the Recall and Precision achieved on a specific topic or category). Topic-specific measures of performance are essential to ensuring that a high level of Recall and Precision is achieved not just overall (where sub-par performance on specific categories of information may not be apparent) but also on the individual component categories of relevance. The ability to obtain topic-specific measures of performance enables H5 to apply its quality control processes at the level of granularity that is appropriate for a given project.
- **Real-time measures of performance.** In order to inform the various decisions that must be made over the course of a document-assessment project, it is necessary that a measurement protocol provide the project team with a view of performance, not just at the project's end, but also at each stage in the assessment effort. H5's measurement protocol provides a project team with real-time measures of performance throughout the course of a project. Equipped with these accurate real-time views of performance, on the one hand, and with the means and expertise to improve performance, on the other, an H5 project team has all the inputs and capabilities required to achieve, predictably and repeatedly, exceptionally high levels of effectiveness.

A measurement protocol is just one component of the overall quality assurance regimen that, from initial reception of data to final delivery of results, guides a document-assessment effort and ensures that all elements are performing as expected; a measurement protocol is, however, the key component of any such regimen. H5's measurement protocol has the characteristics that make it possible both for H5 to perform at an exceptionally high level of effectiveness and for H5 to know that it is performing at that level of effectiveness.

4.2 Implications for confidence in H5's results

H5's measurement protocol also has implications for H5's clients, enabling them both to be confident in the quality of H5's results and to be confident in their ability to demonstrate and defend the quality of those results.

With regard to the quality of results, H5's clients can be certain (a) that H5 centers its QA regimen on the metrics that are the most telling indicators of the performance of a document-assessment process, (b) that H5 uses a sampling design that is the appropriate one by which to obtain estimates of those metrics, and (c) that H5 makes appropriate use of the information provided by its measurement protocol in order to ensure that the output of the H5 document-assessment process is of the expected quality. H5's clients can be confident in the quality of the output of a process that is guided by the measurement protocol we have just reviewed.

With regard to defensibility, H5's clients can be confident that a QA regimen guided by the measurement protocol we have just reviewed will be transparent, easily explained, and readily validated by an independent expert. H5's clients can also be certain that H5, drawing on the information gathered via its measurement protocol over the course of a project, will be able convincingly to demonstrate the quality of its final results.