Measuring Recall in E-Discovery, Part Two: No Easy Answers

John Tredennick

In Part One of this article, I introduced readers to statistical problems inherent in proving the level of recall reached in a Technology Assisted Review (TAR) project. Specifically, I showed that the confidence intervals around an asserted recall percentage could be sufficiently large with typical sample sizes as to undercut the basic assertion used to justify your TAR cutoff.

In our hypothetical example, we had to acknowledge that while our point estimate suggested we had found 75% of the relevant documents in the collection, it was possible that we found only a far lower percentage. For example, with a sample size of 600 documents, the lower bound of our confidence interval was 40%. If we increased the sample size to 2,400 documents, the lower bound only increased to 54%. And, if we upped our sample to 9,500 documents, we got the lower bound to 63%.

Even assuming that 63% as a lower bound is enough, we would have a lot of documents to sample. Using basic assumptions about cost and productivity, we concluded that we might spend 95 hours to review our sample at a cost of about $20,000. If the sample didn't prove out our hoped-for recall level (or if we received more documents to review), we might have to run the sample several times. That is a problem.

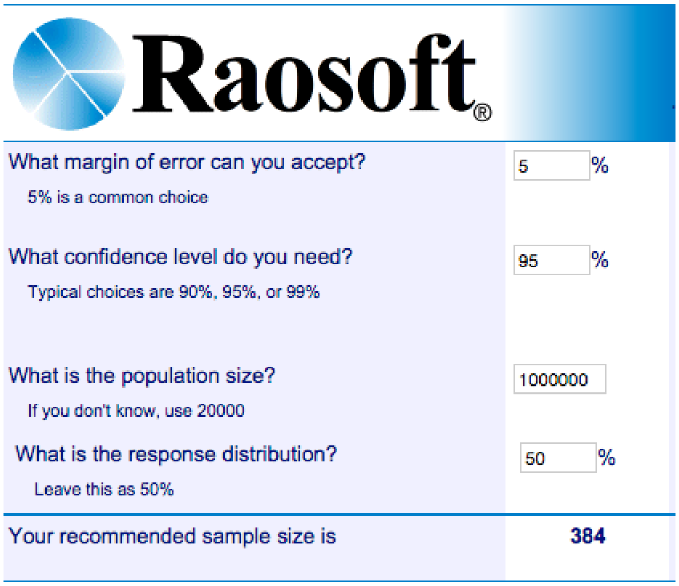
Is there a better and cheaper way to prove recall in a statistically sound manner? In this Part Two, I will take a look at some of the other approaches people have put forward and see how they match up. However, as Maura Grossman and Gordon Cormack warned in "Comments on 'The Implications of Rule 26(g) on the Use of Technology-Assisted Review'" and Bill Dimm amplified in a later post on the subject, there is no free lunch.

The Direct Method

We start with the "Direct Method" for measuring recall. This approach is simple and seems to be universally accepted as a statistically sound method for making this calculation.[1] However, it can be costly to implement.

[1] The Direct Method is mentioned with approval in most of the serious articles I have seen on the subject, including Daubert, Rule 26(g) and the eDiscovery Turkey: Part 2, Tasting the eDiscovery Turkey, by Herb Roitblat; The Implications of Rule 26(g) on the Use of Technology-Assisted Review, by Karl Schieneman and Thomas C. Gricks III; and Comments on "The Implications of

Using the Direct Method, we draw randomly from the entire document population. How many do we draw? That depends on our desired confidence level. For a 95% confidence level, we only need to sample 384 documents.



What margin of error can you accept? 5 %
5% is a common choice

What confidence level do you need? 95 %
Typical choices are 90%, 95%, or 99%

What is the population size? 1000000
If you don't know, use 20000

What is the response distribution? 50 %
Leave this as 50%

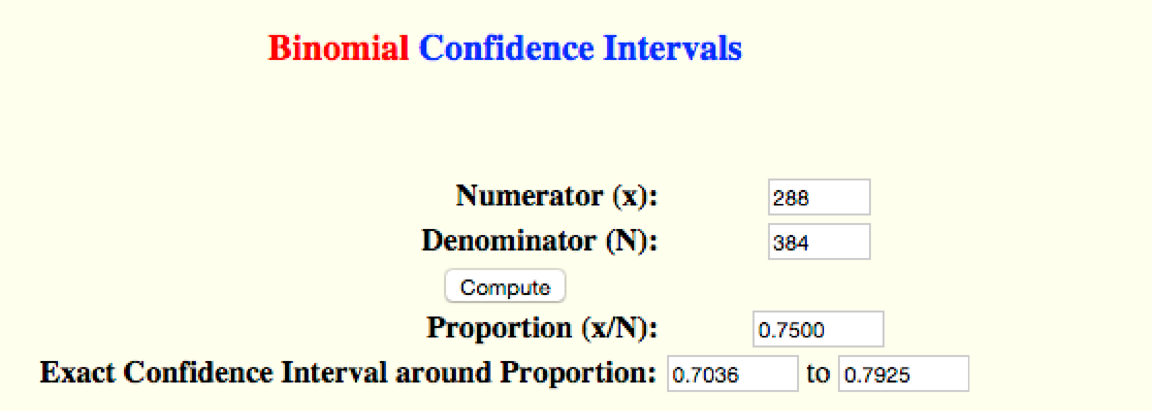
Your recommended sample size is 384

Unfortunately, we are not allowed simply to pick 384 documents at random. Rather, we have to keep picking documents randomly until we find 384 relevant documents.[2] I will talk about why that is a big problem in a minute.

To determine our point estimate, we calculate the percentage of relevant documents that were part of the reviewed set. Let's assume that 288 of the relevant documents were reviewed, leaving 96 of the sampled documents in the discard pile. Using the binomial confidence interval calculator introduced in Part One, we get the following results.

Rule 26(g) on the Use of Technology-Assisted Review," by Maura R. Grossman and Gordan V. Cormack.

[2] William Webber explains that in statistics, the Direct Method is a means to estimate a proportion for a subpopulation, where the (full) population is the collection of documents, the subpopulation are the relevant documents, and the proportion we're interested in is the proportion of relevant documents that are retrieved by the predictive coding system. See Thompson, Steven K., Estimating Proportions, Ratios, and Subpopulation Means. In *Sampling* (pp 57--66), (3rd Edition, 2012) Wiley.



Binomial Confidence Intervals

Numerator (x): 288

Denominator (N): 384

Proportion (x/N): 0.7500

Exact Confidence Interval around Proportion: 0.7036 to 0.7925

In this case, our point estimate suggests we found 75% of the relevant documents, leaving 25% in the discard pile. Our confidence interval suggests that the range of relevant documents is 70-79%, which is a defensible result in my view.

Here's the problem. If the document population is 1% rich, we will have to review on average 100 documents for each relevant one we find. In order to find 384 relevant documents, we will have to look at 38,400 documents on average, and sometimes more. If richness were even lower, say 0.1%, we will have to look at on average 384,000 documents in order to obtain a valid sample.[3] That is a huge burden simply to validate your review results. And, you might have to do it several times before you finish your review.

For collections involving higher richness, for example 10-30%, this approach is less burdensome because it is easier to find relevant examples. As an example, with 30% richness, we might find the desired 384 documents after viewing only a thousand documents. For low richness collections, this approach is a non-starter.

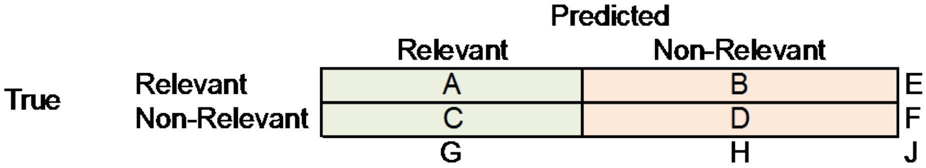
eRecall and the Ratio Methods

Several e-discovery commentators have suggested using ratios to prove recall. All of them work in a similar fashion, comparing point estimates from different parts of the document population—for example an initial richness estimate with

[3] Maura Grossman and Gordon Cormack point out another interesting conundrum stemming from the use of this method. If we found 96 relevant documents from the discard pile, we would have to include them in our production because they are, in fact, relevant. If these documents represented new or different text than documents above the cutoff, one could argue that they should be included as training seeds for a further ranking. They may well elicit even more relevant documents that were similar to the ones adduced from the discard pile in our sample. See M. Grossman and G. Cormack, "Comments on 'The Implications of Rule 26(g) on the Use of Technology-Assisted Review,'" in Federal Courts Law Review, Vol. 7, Issue 1 (2014) at 285, 307.

the richness in the discard pile. For reasons stated in Part One, I don't think any of these approaches work. Why? Because they ignore the impact of confidence intervals.

To understand these methods, we start with what statisticians call a contingency table. It provides a handy way to analyze documents in the review population and to frame certain calculations we might want to make about those documents.



		Predicted		
		Relevant	Non-Relevant	
True	Relevant	A	B	E
	Non-Relevant	C	D	F
		G	H	J

Without trying to cover all the possible combinations, you can quickly see that the documents above a cutoff (predicted relevant) are captured in the green cells A and C (some of which are truly relevant and some of which are not). The total of those documents is represented by item G.

The documents below the cutoff (discard pile) are represented in the red cells B and D (some of which are truly relevant and some of which are not). Their total is represented by item H. E is the total number of relevant documents in the collection (richness). F is the total number of non-relevant documents.

"Elusion" is a term you will hear often in these sorts of discussions. It represents the number of relevant documents that remain in the discard pile (as in eluding your review). In our case, the measure of elusion is B/H (number of relevant in the discard pile divided by the total number in the discard pile).

So, returning to our primary question, how do we determine what percentage of relevant documents we have found and reviewed? Working from our contingency table, we can construct a simple mathematical formula for this calculation:

A (relevant found and reviewed) / E (total relevant)

The answer would be expressed as a percentage, like 75%.

As easy as that formula is to create, it begs the question: How do we prove our recall estimate? In some cases, we know the total for A because we have reviewed all of those documents (prior to review that total would have to be estimated as well). But what is the total for B? That is the difficult question.

Since we can't actually count the number of relevant documents in the discard pile, we are forced to use the same sampling techniques discussed in Part One of this article. And, as I mentioned earlier, we are forced to confront the problem of confidence intervals and their impact on our recall estimate.

eRecall

In his 2013 article, Measurement in eDiscovery, Herbert Roitblat proffers what he calls “eRecall.” In essence, he uses point estimates for initial richness and elusion to determine recall percentages:

> In order to compute recall more precisely, we need to know four things: Prevalence, Elusion, the total number of documents, and the total number of documents designated responsive by process. Two of these (Prevalence and Elusion) need to be estimated from samples. The other two (the number of documents and the number predicted responsive) are known exactly. Knowing these values, we can compute all of the entries in the contingency matrix and can get any of the measures we want with a relatively small investment in documents reviewed.
>
> If we have an estimate of prevalence, either from a random sample drawn before first-pass review or one drawn in the process of the first-pass review . . . , then we have enough information to fill out the contingency table.
>
> If you are familiar with the game Sudoku, then you know about the idea of constraints in a table. In Sudoku, the rows have to add up to a certain number, the columns have to add up, the squares have to add up, etc. Let's start filling out a contingency table in the same way.
>
> First, we know the total number of documents in the collection. Let's say that there are 100,000 documents in the collection (J). We know the total number of documents predicted by the system to be responsive (G); that is the output of our first-pass review. If we know G and J, then we automatically know H, because H is simply J – G. We now have the column totals filled out.
>
> From Prevalence [Richness], we have an estimate of the proportion of documents in the top row, those that are truly responsive. That proportion times the total number of documents is the number of responsive documents estimated in the collection. Again, if we know the total number of documents and the total number of responsive documents, then we know the total number of non-responsive documents. F = J – E. So now, we have the row totals filled out.

> Elusion is used to calculate the number of documents in cells B and D. Cell B is Elusion times the number of documents in the predicted non-responsive column (H). Cell D is the difference between H and B. We now have the second column of the contingency table filled out. Cell A is calculated from the difference between the first row total (E) and cell B, which we just calculated. The final cell, (C) is then the difference between the first column total, G and A. We have now filled out the entire contingency table from knowledge of Elusion, Prevalence, the number of documents emitted by first-pass review, and the number of documents in the collection. We can compute Recall and Precision without a highly burdensome evaluation effort, by sampling a relative small number of documents.

Does this work? Not so far as I can see. The formula relies on the initial point estimate for richness and then a point estimate for elusion. Roitblat suggests that by comparing the two values, we can draw conclusions about the effectiveness of our TAR process. In essence, if the point estimate for richness is 10% and the point estimate for elusion is 0.01%, then the process must have been successful[4].

As I showed in Part One, we could sample the discard pile to estimate elusion (just as we sample the total collection to estimate richness). However, we would still have to take into account the confidence interval around the point estimate. Where richness is low, as in this case, the confidence interval can be wide for large document populations.

Roitblat seems to skirt the issue, at least in this paper. Although he devotes several pages to the topic of confidence intervals, he does not hit the issue head on. Instead, he suggests that we ignore the issue:

[4] William Webber explains that there is a problem trying to draw conclusions from two overlapping samples. Richness, for example is drawn from the entire collection while the elusion sample comes only from the discard pile. This becomes a “live” issue if the two samples say different things about the discard pile. Say the discard sample finds discard richness of 1%, but the part of the richness sample that falls within the discard pile find richness of 2%. We would then have evidence to suggest that the discard sample understates elusion and therefore that eRecall would be an overstatement. See his paper on measuring recall here: William Webber, "Approximate Recall Confidence Intervals", *ACM Transactions on Information Science*, 31:1 (January, 2013), pages 1-33. [http://arxiv.org/abs/1202.2880]
]

> Because the size of the Confidence Interval tells us so little about the quality of results from our review process, there may not be a lot of value derived from selecting narrower Confidence Intervals. One can have very narrow Confidence Intervals around very mediocre performance. Small improvements in the Confidence Interval requires large increases in effort.

With all due respect, I disagree. This is not about the “quality of results” of the review but rather an estimate to support the assertion that we have found a certain percentage of relevant documents. If that assertion is based on sampling, which it must be, you have to specify both the confidence level of your sample and the confidence interval around your point estimate. If the interval is large, which happens with small sample sizes, your recall range can be wide, leaving your production efforts outside the bounds of reasonableness.[5]

Other Ratio Methods

E-discovery veterans Karl Schieneman and Thomas Gricks discuss two other ratio methods in an article appearing in Federal Courts Law Review: The Implications of Rule 26(g) on the Use of Technology-Assisted Review, Vol. 7, Issue 1 (2013). Here is how they described them:

> For example, rather than looking solely at the relevant document population, counsel may choose to calculate recall indirectly by sampling (1) the production, and (2) either the original ESI collection, or the documents identified by the technology-assisted review as not relevant (*i.e.*, the null set). Sampling the production will establish the number of relevant documents produced. Sampling the original ESI collection will establish the total number

[5] Roitblat makes an impassioned defense of eRecall in his blog post, Daubert, Rule 26(g) and the eDiscovery Turkey: Part 2, Tasting the eDiscovery Turkey. He seems to suggest that comparing two sample measures somehow eliminates the problem with confidence intervals. Grossman and Cormack provide test data to suggest that this is not the case. See Maura Grossman and Gordon Cormack, Comments on “The Implications of Rule 26(g) on the Use of Technology-Assisted Review", Federal Courts Law Review, Vol. 7, Issue 1 (2014) 285, 307. In a separate blog posting, Bill Dimm similarly shows that in order for eRecall to be statistically accurate you would have to draw samples roughly as large as those required for the Direct Method, a point also echoed by Grossman and Cormack. See Bill Dimm, eRecall: No Free Lunch, on the Clustify Blog.

> of relevant documents available to be produced; and sampling the null set will establish the number of relevant documents that are not being produced. Recall can then be calculated as the fraction of the total number of relevant documents that are actually being produced in the production set. {citation omitted} This is a more cost effective validation protocol when richness is low, because the size of the samples that must be reviewed is dictated by the confidence criteria, and it is not necessary to review any documents other than those comprising the samples.

The parties reportedly used one of these variants (comparing the production with the discard pile) in the Global Aerospace matter.

Unfortunately, the ratio methods have the same weaknesses as we saw with eRecall. They rely on different point estimates without taking into account the confidence intervals inherent in each estimate. If the samples are small, which was the point of the method, then the confidence intervals will be high. In such a case, the recall estimate is compromised no matter which ratio you use.

Proven Reliability Backed by Post Hoc Sampling

Maura Grossman and Gordon Cormack discuss these issues in their recent publication, "Comments on 'The Implications of Rule 26(g) on the Use of Technology-Assisted Review,'" Federal Courts Law Review, Vol. 7, Issue 1 (2014) at 285. Recognizing the inherent problems with eRecall and the ratio methods, they suggest a wholly different approach to solving the recall proof problem. Their idea is to combine prior successes of the TAR tool being used with limited post hoc sampling. In that regard, they suggest that a Daubert-type analysis[6] could be instructive for this inquiry.

> Notably, the *Daubert* test focuses on *a priori* validation—establishing, through accepted scientific standards, the applicability of the method to the task at hand, its potential error rate, and standards controlling its operation. Therefore, TAR tools and methods should be pre-established as valid in their own right, not re-established *de novo* for each and every matter. (At 311)

They reason that if a particular tool can be shown to be reliable in other TAR projects, it is more likely that it will produce reliable results in the current project.

[6] They carefully and specifically do not propose a formal Daubert hearing, which is required to present certain types of scientific evidence at trial. Rather, they simply suggest that the Daubert approach might be useful for the recall inquiry.

While Grossman and Cormack don’t dismiss the importance of ad hoc testing of the recall results, they suggest that this testing could be less stringent with a system that has already been shown to be reliable.

> Once a sufficient *a priori* confidence level has been established, it should be sufficient to ensure that the method is properly applied by qualified individuals, and that readily observable evidence—both statistical and non-statistical—is consistent with the proper functioning of the method.

I suspect they mean that we could simply rely on the point estimate from a sample of the discard pile and ignore the confidence intervals because we are already confident about our software and techniques.

Does this approach work? I think it might if we could devise some form of testing for the many different TAR algorithms used across the e-discovery industry. To accomplish this, in my opinion at least, we would need to create an independent review body and agree on a method to test different systems. We would also need an agreed-upon set of documents that could be used for this testing.

In the past we have tried to move in this direction through the TREC competitions and more recently through a test sponsored by the Electronic Discovery Institute. The TREC programs proved to be resource intensive and the legal track was discontinued several years ago (although something similar may run in 2015). The EDI project required each participant to review and judge the documents as part of the process, which introduced another form of variability into the testing. In any event, none of these processes are run for the purpose of validating different software methods and algorithms.

Thus, even if we were to agree with the approach (and many do not), I don’t know how we might move forward with a key part of the Grossman-Cormack proposal. Without some form of Underwriters Laboratories inspecting the different TAR systems on the market, how would we ever be able to carry out the suggested approach?

So How Should We Go About Proving Recall?

As you can see, proving the level of relevant documents being produced is a lot more difficult than you might think. If you take the simple approach and use the Direct Method, the process is simple but the effort is great. Sampling tens or hundreds of thousands of documents isn’t practical in most cases. Even sampling 9,500 documents using my technique is a lot of work. But these approaches are statistically sound.

Several of the other methods take less effort in that they don't require large sample sizes. But, they generate wide confidence intervals which may be subject to challenge. If the court will accept a lower bound of 43% or even 50%, then those approaches are fine. Assuming that such ranges will draw objections, we might have to go further.

I like the Grossman Cormack idea of relaxing sampling requirements in cases where the underlying algorithm has been shown to be effective. However, we don't have a neutral body to make those certifications and I don't see that changing soon. So, what to do?

Let me start with the obvious. All of the people mentioned above realize that disproportionate sampling sizes simply run up costs, perhaps with little incremental gain to the parties or the justice system. If the stakes are high and there is reason to doubt the recall point estimate, a sample of 9,500 documents might be justified. In most cases, however, it is not. The point estimate is likely sufficient, especially if a reasonably reliable TAR algorithm and process is used.

Proof Sampling

In an attempt to bridge the gap between the different approaches, here is my suggestion. For starters, I believe the courts should focus on reasonableness over statistical perfection, at least unless circumstances require extraordinary steps. Their inquiry as to the TAR algorithm being used and the process followed should consider its reasonableness rather than some comparative debate over different algorithms and protocols. So long as the algorithm is generally accepted in the industry and a reasonable process followed, that should be enough to move forward.[7]

Then, I suggest the following protocol:

1. For interim sampling (as the review progresses), I would use a small but reasonable sample size, 350 or 600 docs for example, and not worry about confidence intervals. This is merely for internal use to validate the training.

2. When the time comes to sample in support of a final recall assertion, increase the sample size to at least 2,400 documents (95% confidence level and 2% nominal margin of error). We could call this a "Proof

[7] If the algorithm is new and has not been used in the industry, I would require some proof to the effect that it uses a sound mathematical and statistical approach and that a reasonable method was used to employ it. We do not want to shut out new entrants to the market but they should provide some proof of their bona fides.

Sample." I am assuming the point estimate from the Proof Sample backs up the recall assertion you intend to make.

3. I would examine the relevant documents adduced from the Proof Sample set to determine whether they represented important documents (or simply unimportant ones that happen to meet the relevance criteria). If they do, there is a suggestion that the training is still missing key documents and that these new ones should be added to the training as well as the production. A further Proof Sample would be required.

4. When the relevant documents from the Proof Sample do not include any important examples, I would stop the process and be prepared to present the proof to the court and opposing counsel. The level of sampling effort would be deemed reasonable for purposes of the production.

5. To avoid an obvious conflict of interest, the relevant documents from the Proof Sample should be presented to opposing counsel to review. These documents are subject to production anyway; they are, by definition, relevant. Opposing counsel can use them either to accept the Proof Sample or to challenge it. You could also make an argument that opposing counsel should be able to review the sample documents marked irrelevant (assuming they are not privileged) to protect against the possibility that relevant documents are marked irrelevant to ensure that the sample results support the claimed recall.

6. If the opposing party wanted to challenge the proof on the basis of a too-wide confidence interval, put the burden on that party to pay for a larger sample.

7. If that sample showed that in fact the percentage of relevant documents were substantially lower than the proffered level (e.g. 43% rather than 75%), then require the producing party to pick up the costs of the expanded sample and continue training until such time as a reasonable number of relevant documents were produced.

8. The opposing party can continue to challenge the Proof Sample until such time as the recall levels prove to be reasonable.

This challenge idea stems from professional football, which is in full swing as I write this. If a coach feels a call was mistaken, they can issue a challenge. If the call on the field is overruled, there is no penalty for the challenge. If it is affirmed, the coach loses a timeout, which can be important later.

You could bolster the Proof Sample by looking at documents just below the cutoff (in ranked order). This will give you some basis to estimate the cost of finding the next relevant document. That estimate would also bolster the reasonableness arguments one way or the other. Relevant documents found through this sampling could also be presented to support or refute the argument that these documents are not very important to the case.

Does this work? I think it could. While it is not statistically perfect, it is reasonable and it matches a level of effort that is proportionate to its purpose. If opposing counsel feels otherwise, they are free to challenge the assertion through a larger sample.

The goal here is to take advantage of the benefits of TAR, which is a miraculous process that can dramatically reduce discovery costs. I believe every review should use it, particularly if you can take advantage of the flexibility and real-world practicality of continuous active learning. And, I believe that one way or another the courts should agree on a pragmatic way to solve this puzzle that doesn't require reviewing tens of thousands of documents.

If you have a better approach, please share it with us. This is an important issue for our industry.

Postscript #1

Herb Roitblat recently wrote a thoughtful post on his Orcatec blog referencing Part One of my article.[8] Acknowledging the problems inherent in confidence interval analysis, he offered these points about whether we should focus on the statistics of confidence intervals rather than the reasonableness of the process:

> I am definitely in favor of measuring the quality of eDiscovery. I think that it is an important part of establishing trust and of making the case for reasonable enquiry. However, I also think that the goal of eDiscovery is to produce as close to all of the responsive documents as reasonably possible, not to measure it to an arbitrary level of precision. Depending on the prevalence of responsive documents and the desired margin-of-error, the effort needed to

[8] He also shared a draft white paper called "Rational Measurement in Ediscovery," which provides further thoughts on how we should measure the success of the TAR process. He seems to recognize that confidence intervals for the smaller sample sizes represent a problem but makes arguments well worth considering about why that shouldn't be the end of the discussion. Since the paper is not yet posted, I will not comment further on it but it is well worth reading when it comes out.

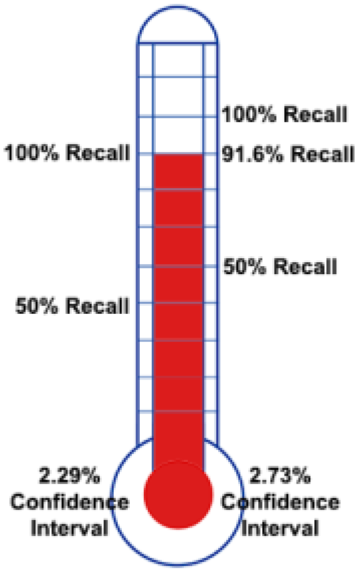
> measure the accuracy of predictive coding can be more than the effort needed to conduct predictive coding.
>
> Until a few years ago, there was basically no effort expended to measure the efficacy of eDiscovery. As computer-assisted review and other technologies became more widespread, an interest in measurement grew, in large part to convince a skeptical audience that these technologies actually worked. Now, I fear, the pendulum has swung too far in the other direction and it seems that measurement has taken over the agenda. Some of the early reported cases involving disputes over the use of predictive coding and some proselytizing by pundits, including probably me, have convinced people that measurement is important. But we risk losing sight of the really important problem, that is good quality eDiscovery.

I think he is right. As I mentioned when I started this series, e-discovery professionals need a reasonable way to measure recall in order to justify review cutoff decisions.

Postscript #2

Just this week, Ralph Losey published a blog post in which he also discusses the challenges in measuring recall. This is the last in a three-part series of posts in which Ralph presents his ideas for visualization of data in predictive coding projects. For visualization of recall, Ralph suggests a thermometer like those used in fundraising drives, but with two different measures. He explains:

> On the left side put the low-end measure, here the 2.29% confidence interval with 229,000 documents, and on the right side, the high measure, 2.73% confidence interval with 273,000 documents. You can thus chart your progress from the two perspectives at once, the low probability error rate, and the high probability error rate. This is shown on the diagram to the right. It shows the metrics of our hypothetical where we have found and confirmed 250,000 relevant documents. That just



> happens to represent 100% recall on the low-end of probability error range using the 2.29% confidence interval. But as explained before, the 250,000 relevant documents found also represents only 91.6% recall on the high-end using the 2.73% confidence interval. You will never really know which is accurate, except that it is safe to bet you have not in fact attained 100% recall.

His post also explains such essential TAR concepts as sampling, probability, prevalence and precision. All three parts are well worth your time to read.