

Faster, better, cheaper legal document review, pipe dream or reality?

Using statistical sampling, quality control and predictive coding to improve accuracy and efficiency

Thomas I. Barnett and Svetlana Godjevac¹

Iron Mountain

Abstract.....	1
Introduction.....	2
Background.....	3
Data Set and experiment	3
Data Set.....	3
Training.....	5
The Task	5
Coding Results.....	5
Analysis	6
Global Agreement Analysis	6
Pair-wise Analysis.....	8
Kappa.....	8
Other Industry Standards	9
Discussion.....	11
Recommendations.....	13
Conclusion	14
References.....	15
APPENDIX.....	16

Abstract

This paper examines coding applied by seven different review groups on the same set of twenty eight thousand documents. The results indicate that the level of agreement between the reviewer groups is much lower than might be suspected based on the general level of confidence on the part of the legal profession in the accuracy and consistency of document review by humans. Each document from a set of twenty eight thousand documents was reviewed for responsiveness, privilege and relevance to specific issues by seven independent review teams. Examination of the seven sets of coding tags for responsiveness revealed an inter-reviewer agreement of 43% for either responsive or non-responsive determinations. The agreement on the responsive determination alone was 9% and on the non-responsive determination was 34% of the total document family count. Pair-wise analysis of the seven groups of reviewers provided higher rates, however no pairing of the teams indicated that there is an unequivocally

¹ Thomas I. Barnett is the leader of the e-Discovery, records and information management consulting division of Iron Mountain, Inc.; Svetlana Godjevac is a senior consultant at Iron Mountain, Inc.

superior assessment of the dataset by any of the teams. This paper considers the ramifications of low agreement of human manual review in the legal domain and the need for industry benchmarks and standards. Suggestions are offered for improving the quality of human manual review using statistical quality control (QC) measures and machine-learning tools for pre-assessment and document categorization.

Introduction

In the world of technology assisted searching, analysis, review and coding of documents in litigation, review by human beings is typically viewed as the gold standard by which the accuracy and reliability of computer designations is measured. Similarly, humans are expected to be able to make judgments with computer-like accuracy and consistency across large sets of data. Expecting computer-like consistency from humans and expecting human-like reasoning from computers is bound to lead to disappointment all the way around. The level of quality of human review of a small number of documents by an expert reviewer familiar with the facts and issues in the matter is in fact a gold standard. But, the typical case involves review of large amounts of data by professional review teams not immersed in the subject matter of the case and the level of accuracy and consistency vary greatly. The levels of accuracy demanded of automated approaches to document classification are expected to confirm to the subject matter expert gold standard not the standard of the typical professional review team. The vast majority of data in legal document review is coded by professional review teams not by the subject matter experts. Thus, holding automated approaches to the gold standard that is barely, if ever, reached in the human review in actual matters creates an unreasonable and likely unachievable goal. This paper proposes that the comparisons be done on a level-playing field and that each approach, human and automated review, be applied to tasks to which they are best suited.

As more human reviewers are applied to the same set of data, the level of consistency and agreement predictably declines. This paper suggests that statistical sampling and statistical quality control is needed to establish a uniform framework from which to assess and compare human and automated review.

The tools used to search, analyze and make determinations about documents in a set of data need to be calibrated and guided by human understanding of the underlying facts and issues in the matter. For now at least, and with acknowledgement of the resounding victory by IBM's *Watson* on *Jeopardy!*, computers don't "understand" things in the way human beings do. Computers can execute vast amounts of simple binary calculations at speeds that are difficult to contemplate. Such calculations can be aggregated and structured in complex ways to mimic human analysis and decision making. But in the end, computers do exactly what they are told and are incapable of independent thought nor can they make decisions outside the scope of their programmatic instructions. Conversely, human beings do not blindly execute precise complex instructions at lightning speed in a predictable and measurable way as computers do. Human creativity and independent thought result in variability and unpredictability when attempting to make large numbers of fine distinctions. The independence and creativity that allows a person to make a novel observation or discovery is the flip side of the lack of the ability to make fast, mechanically precise consistent determinations about documents. This paper proposes considering a set of documents for review in a litigation as a continuum of relevance to a set of criteria rather than as a set of uniform discreet yes/no determinations. Under that model, the review process can be designed to play to the relative strengths of computer and human analysis. Within any typical set of data, certain documents will be clearly responsive. Others will be clearly non-responsive. The remaining documents can be characterized as having an ambiguous classification. Trying to get computers to accurately assess documents that humans find ambiguous is not effective—it plays to the computer's weakness. Computers should be utilized where they are strongest—quick, fast, accurate determinations of clear cut binary determinations. By contrast, for documents that are not clearly responsive or non-responsive, human judgment, creativity and flexibility is best suited to make the judgment calls. Based on this model, this

paper asserts that computers should be used to classify non-ambiguous documents while human reviewers should focus attention on documents whose classification is ambiguous.

This paper examines coding applied by seven different review groups on the same set of twenty eight thousand documents. The results indicate that the level of agreement between the reviewer groups is much lower than might be suspected based on the general level of confidence on the part of the legal profession in the accuracy and consistency of document review by humans (see Grossman and Cormack, 2011 for a similar position). However, a comparison to other industries, such as medical text coding for example, suggests that the legal industry is on a par with the results in other industries. This should not be surprising considering that both tasks are language-based tasks involving interpretation and translation of vast amounts of text into a single numeric code. This paper argues that the identified distribution of disagreements among human reviewers suggests that the nature of the task itself will never allow significant improvement in human review without disproportionate additional cost and time spend reviewing and cross checking document determinations. A proposed method to achieve higher consistency and accuracy lies in redistribution of the task between humans and computers. Computers should be allowed to jump-start the review, as they will easily recognize high-certainty sets, and humans should focus on ambiguous, middle of the scale sets, as only human analytical and inferential ability can successfully classify the documents of ambiguous classification.

Background

This experiment was originally conducted as a pilot by a company for the purpose of selecting a provider of document review services. The intent was to compare the document coding of five different document review providers against a control set of the same documents coded by outside counsel. The results of the six team review (five document review vendors and the outside counsel team) proved inconclusive to client in determining which provider to select. Subsequently, the client decided to assess the quality and accuracy of the providers' coding of the documents using the assessments of a different outside counsel who had reviewed the same set of documents. This second control group constituted the seventh set of human manual assessments for each document in this set. The additional control group's document coding determinations were ultimately not considered definitive and the pilot did not result in any clear "winner."

The analysis was performed on the final aggregate set of document coding from all review teams and does not assume that the coding of any one group is the ground truth. The client concluded that neither of the two control groups was able to provide coding that was of sufficient accuracy to be considered a gold standard. From the client's perspective, the experiment failed, as it was not possible to determine a winner among the document review service providers. Nevertheless for purposes of this analysis, the data provided a unique and valuable source of information for the eDiscovery industry and it is hoped that the results can be instructive in conducting comparisons of document review groups as well as creating quality control standards and workflow improvements for legal document review.

Data Set and experiment

Data Set

The reviewed document population for this experiment consisted of a sample of the electronically stored information (ESI) from six different custodians. The starting set contains 12,272 families comprised of 28,209 documents. Of the total 28,209 documents, most of the documents were emails and Microsoft Office application files. The basic data composition is represented in Figure 1. The most common family

unit² size was two. The majority of the corpus, 99%, consisted of families with no more than eight attachments. The family size frequencies are provided in Figure 2.

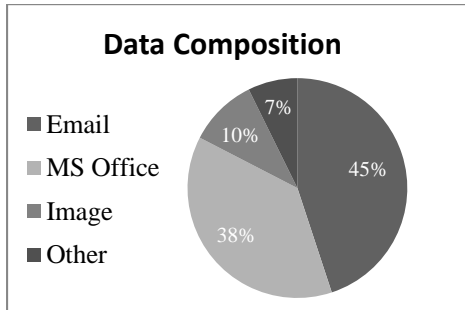


FIGURE 1- DATA COMPOSITION OF THE REVIEW SET

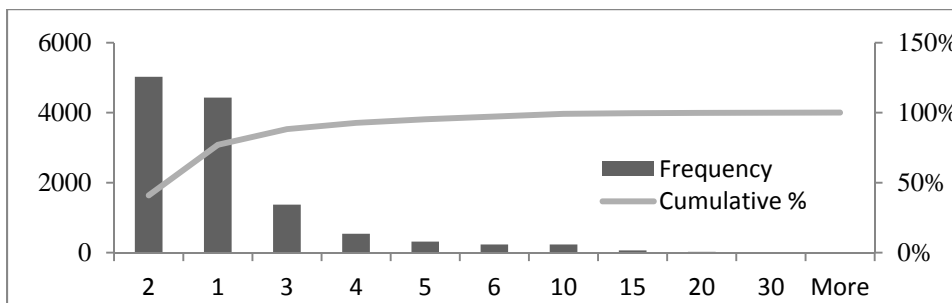


FIGURE 2 – FREQUENCY DISTRIBUTION OF FAMILY-UNIT SIZE – Most families consisted of two or one member.

Bin	Frequency	Cumulative %
2	5023	40.93%
1	4432	77.05%
3	1375	88.25%
4	542	92.67%
5	318	95.26%
6	235	97.17%
7-10	233	99.07%
11-15	66	99.61%
16-20	25	99.81%
21-30	13	99.92%
31 or More	10	100.00%

TABLE 1 – HISTOGRAM TABLE FOR THE FREQUENCY OF DISTRIBUTION OF SIZE OF FAMILY-UNITS

Due to errors in coding, the original set had to be cleaned up for the purpose of analysis. Forty-seven document families were excluded because at least one member has been coded “Technical Issue.” Ninety five families were excluded because one or more members in the family were not coded consistently with the rest of the family. A summary of the data exclusion is presented in Table 2.

	ORIGINAL	EXCLUDED TECH ERRORS FAMILIES	EXCLUDED INCONSISTENT FAMILIES	CONSISTENT FAMILIES FINAL COUNT
Documents	28,209	205	350	27,654
Families	12,272	47	95	12,130

TABLE 2 – DATA SETS THAT WERE EXCLUDED FROM THE ORIGINAL SET AND THE FINAL SET COUNTS

² A “family unit” for purposes of this paper means an email and any associated attachments.

Reviewers

Seven reviewer groups were provided with access to the data for assessment. The review was conducted by groups of attorneys employed by five different legal document review providers and groups of litigators at two different law firms. Each group had a range of between six and seventeen attorneys who were provided access to the data.

Training

Each reviewer group received approximately three hours of subject matter training by the first law firm and the client. They were also provided with a review protocol, a coding manual, and an hour of training on the review platform. Each reviewer also received a binder with the review protocol, the official complaint, a list of acronyms and other subject matter materials necessary for document assessment. All but one team used the same hosted review platform which they accessed in a controlled environment during business hours. One group, group F, performed the review on their own platform, although there is no data to suggest that that influenced the document coding decisions.

The Task

The documents were arranged into batches of approximately 100 (keeping family units together). The batches were made up of randomly selected document families from the data set. The task involved reviewing and coding each document in the batch before the next batch could be requested. The coding tags included assessments for responsiveness, privilege, issue, and “hot” (significant) document designations. The assessments were made at the family unit level rather than by the individual component of a message unit. For example, if any member of the family was considered responsive, the entire family was coded responsive. Similarly, if any member of the responsive family was considered privileged, the entire family was tagged privileged. Each review team performed quality control checks according to their standard practice before providing the coded documents to the client.

Reviewers also had an option to tag documents for any technical problems, such as difficulty in viewing or errors in processing. Some of these errors prevented reviewers from making assessments for responsiveness and privilege. Consequently, due to the absence of coding for responsiveness, 205 documents were excluded from the overall agreement comparisons.

For purposes of analysis, responsiveness determinations were the sole focus. Unlike issue coding, these assessments are binary and all documents must be coded either responsive or non-responsive. Privilege determinations were not included because the privilege rates were very low, less than 1%, and were dependent on the responsive assessment (i.e., if a document was coded non-responsive, no determination would be made as to whether or not it was privileged).

Coding Results

The responsiveness rates among the seven review groups range from 23% to 54% of the total families. The difference spans 31% with a standard deviation of 0.11. The coding of each review group is presented in Table 3 below.

Tag Count per Family	Group						
	A	B	C	D	E	F	G
Non-Responsive	8279	5560	7641	9331	8842	6054	7316
Responsive	3851	6570	4489	2799	3288	6076	4814
Total	12130	12130	12130	12130	12130	12130	12130
Responsive Rate	31.75%	54.16%	37.01%	23.08%	27.11%	50.09%	39.69%

TABLE 3 – CODING COUNTS FOR EACH REVIEW TEAM

By definition, the global inter-reviewer agreement (the percentage of document coding all groups agree on) cannot exceed the lowest responsiveness rates found among all seven groups. In other words, the maximum rate of agreement cannot be higher than the sum of the lowest proportion of responsive tags among all the teams and the lowest proportion of non-responsive tags among all the teams (i.e., $23.08\% + 45.84\% = 68.92\%$).

Analysis

Two types of analyses were conducted: a global analysis of agreement, and a pair-wise agreement analysis. In the global analysis, the level of agreement between all reviewer groups was the focus. Sets of documents on which different teams agreed were identified: a set of documents for which all seven groups agreed, or 7/7, sets of documents for which six out of the seven agreed, or 6/7, five out of seven, 5/7, and four out of seven, 4/7. The remaining combinations are the inverse of these four. The pair-wise analysis was performed in two ways: an agreement expressed as a percent overlap between a pair of review teams and agreement expressed as Cohen’s Kappa coefficient.

Global Agreement Analysis

The analyzed document set had 12,130 family units with a total of 27,654 documents. The set of families for which all seven groups agreed on responsiveness (either the responsive or non-responsive tag), is 5,233 family units, or 43.14% of the data set. Six groups agreed on 2,482 family units, or 20.46% of the data. Five groups agreed on 2,120 family units, or 17.48% of the data and four groups agreed on 2,295 family units, or 18.92% of the data. The agreement results are shown in Figure 3.

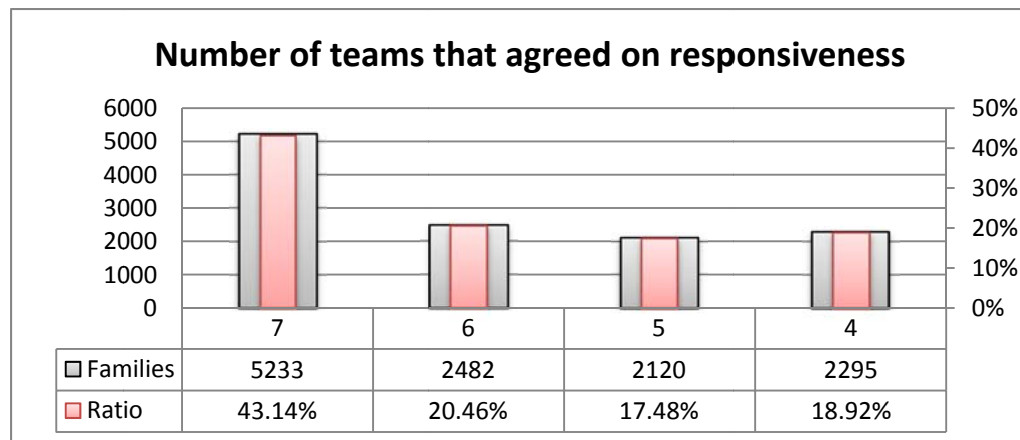


FIGURE 3 – REVIEWER AGREEMENTS ON RESPONSIVENESS

The chart shows the number of document families and the number of teams that tagged the documents the same way. For example, all seven teams coded 5233 families the same way.

The data in Figure 3 include agreements on both responsive and non-responsive determinations. Breaking down this agreement into its constituent parts and considering only the responsive tag (the non-responsive tag is a mirror image of the responsive tag) shows that the reviewers agreed more often on non-responsive than on the responsive tags. The distribution of the responsive tag agreement is provided in Figure 4.

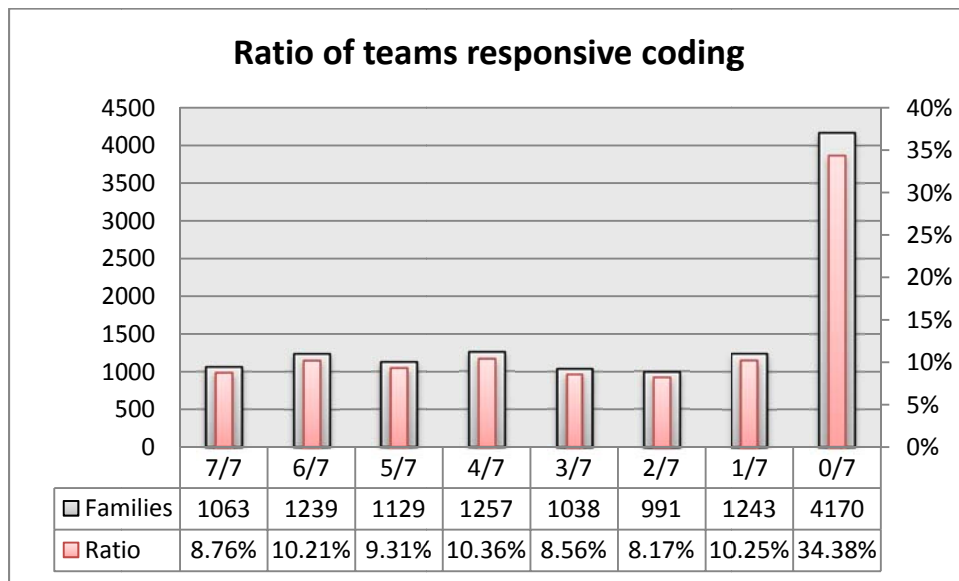


FIGURE 4 – DISTRIBUTION OF THE RESPONSIVE TAG ACROSS REVIEW GROUPS

The chart shows how many document families different number of teams' coded responsive. For example, seven review teams agreed on 1063 families being responsive; 6 review teams agreed on 1239 families being responsive etc. No group agreed that 4170 families were responsive, i.e., all seven groups coded this set as non-responsive.

All-groups agreement of 43.14%, shown in Figure 3, is the sum of the 8.76% document pool for which all seven teams said the document family was responsive and the 34.38% document pool for which all seven teams said the document family was non-responsive. The higher non-responsive agreement can be viewed as a result of the low responsive rate found in most groups coding. Two out of seven reviewer groups had responsive rates less than 30% (groups D and E, see Table 2).³ With fewer documents coded responsive by two groups, the overall agreement for responsive would not be expected to be higher than the lowest responsive rate (Group D). Similar reasoning can be applied to the non-responsive rates.

If the seven groups applied the coding simply by guessing, what level agreement would be expected? With the seven review teams, a pure guessing approach would be the equivalent of seven coin tosses for each family of documents. Each coin toss is a binary decision, heads or tails, similar to document responsiveness tagging. The probability of having a chance agreement by having reviewers guess rather than apply analysis and reasoning among the seven groups is 1.56%, or $2/128$.⁴ The achieved 43% agreement is thus evidence of a decision and not a mere guess. However, the decision would require more specificity if it were to be executed in perfect accordance at a higher frequency than 43%. The next consideration was pair-wise comparisons⁵ and the level of correlation among the group pairs.

³ This is evidence that the actual, though unknown, number of responsive families, is much lower than the number of the actual, also unknown, number of non-responsive families.

⁴ The probability for seven out of seven agreement on each responsive or non-responsive is $\left(\frac{1}{2}\right)^7$, which is $\left(\frac{1}{128}\right)$.

Since aggregate agreement was computed, i.e., including agreement on both responsive and non-responsive those two probabilities are added to arrive at $\left(\frac{2}{128}\right)$.

⁵ Pair-wise comparison is a common scientific method for calculating a relationship between a pair of results to determine which member of the pair is better or has a greater level of property that is under discussion.

Pair-wise Analysis

This analysis presents calculations of percent overlap (or agreement) between any two groups. The results are given in Table 4. Overlap is defined as the sum of all document families where two review teams agreed in responsiveness (responsive and non-responsive tag agreement) divided by the total number of document families they reviewed. The raw agreement values are shown in Table 9 in the Appendix.⁶

	A	B	C	D	E	F	G
A							
B	75.06%						
C	83.05%	75.01%					
D	74.51%	65.53%	72.20%				
E	79.91%	71.95%	76.69%	80.32%			
F	76.94%	84.90%	75.21%	68.17%	74.26%		
G	76.94%	75.23%	74.11%	67.39%	73.08%	77.20%	

TABLE 4 – PAIR-WISE AGREEMENTS

The table presents percent overlap of tagging assessments between a pair of review teams. For example, A and B teams tagging overlapped 75% of the time.

The highest overlap was achieved by groups A&C (83%) and B&F (85%). The lowest overlap was manifest between groups B&D (66%). The average overlap between group pairs is 75%. The group average aligns very closely with the results from a recent study by Roitblat et al.(2010) that compared agreement of pairs of manual review teams. Their comparison of manual review indicated that two different human review teams agreed with the original assessment at remarkably similar levels to the ones presented here. Their Team A agreed with the original review 75.58%, and Team B agreed with the original review 72.00%. So, results presented here replicate and reinforce the results presented in Roitblat et al. (2010). However, an earlier TREC study (Voorhees 2000) provided much lower agreement levels. In that study three different pairs of manual review teams had overlaps of 42.1%, 49.4% and 42.6%. It is not clear though, how the difference in ~30% agreement between the more recent studies and Voorhees’ might be accounted for.

The average 75% coding overlap between two review teams suggests that even among the professional reviewers one in every four documents is not agreed upon. This result challenges the common assumption that there are discernable right and wrong determination for every document and that such a determination will be reached uniformly by different human reviewers.

Kappa

To further examine the level of agreement of responsiveness tagging between reviewer groups, Cohen’s Kappa coefficient was computed. The Kappa coefficient is a measure of a level of agreement between two judges on a sorting of any number of items into a defined number of mutually exclusive categories. In our scenario, each review team is a judge and responsiveness tagging is a sorting into two mutually exclusive categories (responsive and non-responsive). Kappa coefficient values can range between 1 (complete agreement, or far more than expected by chance) to -1 (complete disagreement, or far less than expected by chance), with 0 being a neutral case, or as one would expect by pure chance. This coefficient is regarded as a better measure of agreement than percent-overlap because it eliminates the level of chance-agreement from its value. Landis and Koch (1977) propose the following interpretation of Kappa scores:

⁶ Overlap presented in Table 4 was calculated from the values provided in Table 9. A and B teams agreed on 3698 document families being responsive and 5407 document families being non-responsive. Their coding then overlapped 75.06% $((3698+5407)/12130=0.7506)$.

- 0.01-0.20 – Slight agreement
- 0.21-0.40 – Fair agreement
- 0.41-0.60 – Moderate agreement
- 0.61-0.80 – Substantial agreement
- 0.81-0.99 – Almost perfect agreement

Kappa values for the seven review groups are presented in Table 5. Using the Landis and Koch interpretation scale for the Kappa scores, most of the team pairs, 13 of them, show moderate agreement. Their Kappa values range from 0.45 to 0.54. Two team pairs show substantial agreement, and six team pairs show fair agreement. The lowest score is 0.3402 (Groups B & D), and 0.6979 is the highest (Groups B & F). The Kappa values confirm the pair-wise analysis of percent-overlap for the groups: B&F exhibit the highest overlap and B&D the lowest on both analyses.

	A	B	C	D	E	F	G
A							
B	0.5159						
C	0.6255	0.5108					
D	0.3655	0.3402	0.3536				
E	0.5175	0.4597	0.4709	0.4776			
F	0.494	0.6979	0.5044	0.364	0.4857		
G	0.5013	0.5131	0.4528	0.4053	0.4053	0.5441	

TABLE 5 – KAPPA COEFFICIENT

The Kappa scores range [0.3402 - 0.6979] is similar to the one found by Wang & Soergel (2010) in their study of inter-rater agreement between two groups of human reviewers. Their experiment involved four law students as the LAW team and four library and information studies students, as the LIS team. The goal of their experiment was to test whether the legal background affects the quality of document review. The Kappa mean scores within the LAW team, within the LIS team and across the two teams show remarkably similar ranges: (a) within LAW [0.38 – 0.69], (b) within LIS [0.30 – 0.54] and (c) across LAW and LIS [0.47 – 0.61]. The range of the Kappa coefficient for Wang and Soergel’s LAW group closely parallels the range reported here for the seven review teams.

The Kappa coefficient analysis further confirms that humans reviewing the same documents frequently disagree. As discussed below, this fact suggests that greater focus on quality control is warranted.

Other Industry Standards

In order to put results presented here into a broader context, a short overview of similar tasks in other domains is presented. There are a variety of applications that require translation of natural language into other systems, whether other natural languages or man-made systems. Document review coding is an example of a man-made system that requires a translation from document text into review codes. Tasks of this nature could theoretically be automated if explicit sets of rules could accurately be defined in advance. For tasks that involve natural language, the number of explicit rules is too numerous to be able to be defined in advance. One solution to this problem is machine learning. Machine learning is a substitute for pre-defined set of rules. In the absence of explicit rules, a machine learning program uses input from a training set and “learns” how to apply it in situations that are similar to the ones in the training set. Machine learning is used in search engines, natural language processing, detecting credit card fraud, stock market analysis, handwriting recognition, game playing, medicine, and many others areas.

The training phase of machine learning requires high quality human input, where the high level of accuracy is confirmed through agreement with multiple human experts on the same task.

The medical industry has been faced with the challenge of coding millions of records for medical diagnosis, billing and insurance purposes, among others. In the domain of patient records, a medical diagnosis is required to be translated into a billing code. The billing codes are based on the classification provided by the World Health Organization in the International Classification of Diseases (ICD). The process of human coding of medical diagnoses is challenged by the existence of thousands of possible codes, which is both time-consuming and error-prone. To alleviate the burden and improve consistency of human coding, a number of machine learning systems for classifying text using natural language processing have been designed and implemented in the medical industry. The “training” of the system, using a set of documents that have been coded by highly trained human ICD coding experts is critical to the accuracy of all of the ICD automated coding systems.

The application of ICD codes for medical diagnosis is in many ways similar to legal document review. Both involve reading and understanding natural language texts (or listening to audio files) and applying a code as an output of the process. The ICD codes are directly parallel with issue coding in legal document review in that a number of possibilities per document are open for assignment. The interpretation of natural language (verbal encoding of someone else’s intentions) is at the core of the process in both tasks. Responsive and privilege binary distinctions are a simpler form of coding than relevance to a specific issue in a lawsuit as the number of possibilities are reduced to two. So, the agreement results achieved in responsiveness tagging are expected to be higher than agreements on issue tagging in legal review or ICD coding in medical review due to the smaller number of choices a reviewer/coder is faced with.

The literature on training and automation of the ICD coding assignment and other systems for classification of medical information, such as SNOMED (Systematized Nomenclature of Medicine), is vast. Kappa is often used as a measure of inter-reviewer agreement and for comparison of automated system against human review, the most commonly used metric is the harmonic mean of precision and recall, or the F-score.⁷ This score can only be computed if precision and recall can be computed. Having a gold standard is the key to all machine learning systems as well as the evaluation metrics. If the “true” answer is unavailable, the system is unable to learn.⁸ Some examples of results in the medical domain are provided below.

Uzuner et al. (2008) measured inter-annotator agreement of the patient’s smoking status based on the hospital discharge summary. The annotators were two pulmonologists who provided annotations relying on the explicit text in the summary as well as their understanding of the same text. The metric shown in Table 6 is the Kappa coefficient. The intuitive judgment values are the most directly comparable to the document review assessments as they rely on human ability for interpretation. These scores are similar to the ones reported here for attorney teams. The overall range is wider with the highest score in the “almost perfect” category.

⁷ The F-score is computed as $2 * P * R / (P + R)$, where P is precision and R is recall. Precision is a metric that quantifies how many of the retrieved documents are correct and precision is a metric that quantifies how many correct documents were missed. In order to calculate these values, the number of correct documents must be known. The set of correct documents is what is referred to as the “gold standard.”

⁸ Human intelligence, although incomparably more flexible and dynamic in comparison to a machine, is also dependent on the “system updates”, or the feedback loop for arriving at the truth. Quality control checks of a sample of documents being reviewed often serve to provide feedback to the reviewers on the accuracy of their coding choices so that they can make course-corrections going forward. This process is an important calibration tool in manual review.

Agreement	Textual Judgment	Intuitive Judgment
Observed	0.93	0.73
Specific (Past Smoker)	0.85	0.56
Specific (Current Smoker)	0.72	0.44
Specific (Smoker)	0.40	0.30
Specific (Non-Smoker)	0.95	0.60
Specific (Unknown)	0.98	0.84

TABLE 6- KAPPA COEFFICIENTS FOR INTER-ANNOTATOR AGREEMENT FOR PATIENT'S SMOKING STATUS From Uzun et al. 2008 study on patient smoking status from medical discharge summaries. The study shows the kappa scores for assessments based on explicit text and interpretive judgments based on human understanding.

Table 7, below, shows pair-wise comparison of inter-reviewer agreement using the F-measure, for three human annotators for ICD-9-CM codes applied to radiology reports on a test set (unseen data). The F-scores of the training set were approximately 2 points higher in each case. This higher measure is as one would expect, as the training set is the set that they've seen prior to evaluation.

	A1	A2	A3
A1		73.97	65.61
A2	73.97		70.89
A3	65.61	70.89	

TABLE 7- INTER-ANNOTATOR AGREEMENT ON ICD-9-CM CODING OF RADIOLOGY REPORTS (Richard Farkas And Gyorgy Szarvas, 2008)

Crammer et al. (2007) study of inter-annotator agreement for ICD-9-CM coding of free text radiology reports, also using three human coders, the average F-measure of 74.85 (with standard deviation of 0.06). Resnik et al. (2006) provide measures of inter-annotator agreement on task involving code application for ICD-9-CM and CPT (Current Procedural Terminology) on a random sample of 720 radiology notes from a single week from a large teaching hospital. Their evaluations show averages for all annotators. They've used a proportion measure for ICD. Their results are provided in Table 8.

	ICD
Intra-coder agreement	64%
Inter-coder agreement	47%

TABLE 8 – INTER AND INTRA CODER AGREEMENT ON ICD CODE ASSIGNMENTS (Resnik et al. 2006)

In all of the radiology coding tasks presented above, the inter-reviewer agreement is not dramatically different from the agreements found in this study of legal document review. Given that similarity of tasks, this suggests that manual (human) review of discovery documents should not be expected to improve significantly unless additional means are used to help better allocate time for human review of more complex documents that need to be assessed with more attention.

One common thread to the medical studies and the studies on legal document review, whether by humans or machines, referenced here is the fact that none of them show results approaching full agreement or high retrieval (measured by the F-score). Both fields appear to be at the same level of advancement when it comes to coping with the inherent ambiguity of human language.

Discussion

Results and their implication

The global agreement calculations show that reviewers unanimously agreed on nearly half the documents, or 43%. This set of documents can be termed a high certainty set. On the other roughly half of the

documents, the reviewers had varying degrees of certainty, 6/7, 5/7, and 4/7. This distribution of varying degrees of collective uncertainty can be viewed as a consequence of the “translation” reviewers had to make in order to force a simple yes/no determination onto intrinsically subjective nonlinear data. In other words, the perspective of multiple review groups reviewing the same set of documents rather than a single review team provides support for the intuitive understanding that documents have varying degrees of relevance. When reviewers are asked to code documents either responsive or non-responsive, they are essentially being asked to translate a continuum of degrees of responsiveness into a threshold that will create a single artificial boundary for a yes/no determination. Where this boundary lies is subject to interpretation. The subject matter training the reviewers receive at the beginning of a review is supposed to train them to find this boundary uniformly at the same place every time. However, in reality, each reviewer (and consequently each group) arrives at a different threshold that defines that boundary. Quality control is needed to moderate the understanding of the boundary placement throughout the review. The level of QC needed to guarantee that this boundary is perfectly calibrated and aligned for all reviewers is not practical in terms of time and cost in the context of legal document review.

Part of the quality of control process in the context of document review is evaluation of performance. The most effective means of evaluating quality of performance is to use a quantifiable system. Often used steps for quantifiable evaluation of language-based tasks are:

- a) comparison to a gold standard
- b) inter-coder agreement (consistency across multiple reviewers)
- c) intra-coder agreement (consistency within the same reviewer)

This study of agreement only focused on inter-coder agreement. Access to a gold standard was unavailable and inter-reviewer consistency either at the group-level or reviewer-level would require more complex computations such as creating document sub-groupings based on content similarity and assessing consistency of coding within each subgroup within a reviewer, within a reviewer team and across all reviewer teams.

Comparison of inter-reviewer agreement from these seven groups to the quoted radiology annotators shows that the legal review groups are on a par with the medical profession. The ICD proportion for inter-coder agreement was 47% (Table 8). This value is directly comparable to the average value of 75%, calculated in Table 5 for legal document review. The comparison of these two values gives legal review a superior grade. The comparison analysis, however, must acknowledge that the ICD coders use thousands of codes, rather the just two (i.e., responsive or non-responsive), as do the legal document reviewers and thus the probability of agreement is reduced by the larger number of possible choices.

If it is assumed that the set of varying degrees of certainty (the sets where agreements were 6/7, 5/7, and 4/7) and the sets outside of agreement (intersections) in the pair-wise comparisons are the sets that contain errors, the nature of these errors and the cost associated with them needs to be considered.

Error types and their cost

Errors are divided into two types:

- False positives (Type I error) – documents coded responsive, but are actually non-responsive.
- False negatives (Type II error) – documents coded non-responsive, but are actually responsive.

False positives are typically caught by QC and/or additional review passes. This is because the set of responsive documents is usually further reviewed either for assessment/confirmation of privilege, privilege type or redaction. Errors of this type, Type I, are usually more costly for the client in the field of legal document review, because these types of errors may result in waiver of privilege or revealing potentially damaging information to the opposing side.

False negatives and the degree of their presence in the non-responsive set usually remain undiscovered, unless active measures are taken to identify them such as re-review or inferential statistics through sampling.. This type of error is often neglected as it is less costly from the perspective of the risk of unintentional information exposure. However, if detected by the opposing side, it could lead to sanctions for withholding relevant information.

In this study, an assumption was made that the gold standard for this set was not available, However, if the set of 7/7 agreements for responsive and non-responsive were to be used as the gold standard, the calculation based on this gold standard would be biased in favor of the groups who made conservative judgments on responsiveness. So, this evaluation cannot be used as a measure of quality of the review groups, although it could be used as a way of measuring the cost of error for the client.

Recommendations

Sharing the work

The distribution of partial agreements, viewed as a continuum of degrees of certainty, is analogous to the predictive coding systems whose output is a probability score for each document, rather than a binary decision on category membership. If human review manifests a continuum of certainty levels with respect to relevancy judgment anyway, why not then share the task of review with the predictive coding systems which automatically output degrees of certainty?

Sharing the task does not mean fully delegating, but rather incorporating predictive coding technologies to aid human document review by using computer software to segregate the high-certainty sets (the high probabilities and the low probabilities for category membership, or the 7/7 and 0/7 agreements in this study) and allow human experts to focus on the middle range probabilities (the 6/7, 5/7, and 4/7 in this study). The high certainty sets are the easy calls to make as they are more clear-cut and so they should be delegated to the low cost (computer) labor. The difficult decisions are the decisions that require human intelligence for disambiguation as well as strong subject matter expertise.

The generated probabilities can also speed up the review of the middle of the scale sets. Resnik et al. (2006) show that computer assisted workflow improves human scores by 6% in ICD coding. This improvement in speed may come with a bias, however, and so, it should be considered carefully. They note that:

“Post hoc reviews can overestimate levels of agreement when complex or subjective judgments are involved, since it is more likely that a reviewer will approve of a choice than it is that they would have made exactly the same choice independently”

Whether predictive coding should be revealed to the reviewers for the middle of the scale sets is a decision that will require determination on a case-by-case basis.

Feedback

Feedback is essential for any learning environment. Legal document review is a business process that starts anew with each case. The task begins typically after no more than a day of training, if that. Due to the high costs of document review by attorneys, the learning phase is becoming shorter and shorter and the expectation is that even very complex subject matters can be absorbed in short time frames. Unfortunately, that assumption is to the detriment of the depth of expertise reviewers can attain and consequently the quality of the review. The actual subject matter experts rarely review documents and thus the true gold standard is an illusion. To improve the quality of review, continuous dynamic updates of expert judgments provided to the reviewers are critical. If reviewers receive feedback about the

accuracy of their work promptly, fewer errors will ensue. This result will minimize the need for recoding after quality control checks are performed as fewer errors should be present.

Statistical QC

Current legal document review practices rely more often on judgmental sampling as a QC procedure than on statistical sampling. Although judgmental sampling has value in the QC process, it also has deficiencies. The key detriment is the inability to apply inferences to the larger set. So, while judgmental sampling may reveal errors, there is no way of estimating if the types of errors the QC team didn't consider are present and the degree to which they may be present in the population as a whole. For example, because judgmental sampling deals with the known risks the searches target known "keywords" to create samples for QC. The end result is that unanticipated uses of language to describe the high-risk activities at the core of review will remain undetected. Implementing statistical sampling for the QC process would allow document review to provide quantifiable metrics on the quality of the output and it would also create a higher chance of finding unanticipated references that may inform new searches and require document recoding.

As predictive coding is becoming a more widely available offering in the practice of legal document review, it is essential that the double standard that seems to be applied to this programmatic approach as compared to the standards for human review be addressed. Clients uniformly require that predictive coding come with 95%-99% accuracy. This level of accuracy for the machine is expected because the assumption is that human review is in the 100% range of accuracy (for a similar discussion see Grossman and Cormack 2011). There are at least two problems with this reasoning. First, no research was uncovered that suggests that human accuracy level ever approaches 100% accuracy. Second, it seems that this unsupported assumption is also tacitly known to be false. Either way, the predictive coding should be welcomed by the legal community and judged by the same, not higher, standards than manual review. In order to provide the ground for comparison and equivalent standards of quality, manual review should incorporate statistical QC into its workflow as only with this type of quality check can measures of accuracy, such as precision and recall, be calculated.

Conclusion

Document review for litigation discovery is demanding, time-consuming, expensive and risky. It requires both the ability to perform routine repetitive tasks in an accurate and timely manner as well as the ability to apply human judgment, reasoning and making fine distinctions about complex matters. And the faulty decisions can have tremendous legal and financial consequences. Neither humans nor computers are perfectly suited to accomplish these diverse tasks. The recommended approach to achieve greater accuracy and efficiency is to allocate tasks between humans and computers that play to their respective strengths rather than to their respective weaknesses. Computers perform high speed, repetitive tasks far more efficiently than humans. But computers have no ability to use reason, creativity or judgment beyond the predefined rule sets that are used to program them. Large sets of documents subject to review in litigation contain a continuum of responsiveness. That is, there are some documents that are clearly responsive, some that are clearly non-responsive and the remainder are somewhere in between. Efficiency and accuracy in legal document review can be improved by allocating computer assisted sorting and categorization processes to the high certainty ends of the continuum while human reviewers focus their time and attention using their uniquely human analytical and inferential ability classifying the ambiguous documents.

References

- Richard Farkas and Gyorgy Szarvas, “Automatic construction of rule-based ICD-9-CM coding systems”, *BMC Bioinformatics* 2008, 9.
- Koby Crammer and Mark Dredze and Kuzman Ganchev and Partha Partim Talukdar, “Automatic Code Assignment To Medical Text”, *BioNLP '07 Proceedings of the Workshop on BioNLP 2007: Biological, Translational, and Clinical Language Processing* 2007.
- Maura R. Grossman & Gordon V. Cormack, *Technology-Assisted Review in E-Discovery Can Be More Effective and More Efficient Than Exhaustive Manual Review*, XVII RICH. J.L. & TECH. 11 (2011), <http://jolt.richmond.edu/v17i3/article11.pdf>
- Landis, J.R. and Koch, G. G., “The measurement of observer agreement for categorical data”, *Biometrics* 1977, 33.
- Philip Resnik, Micahel Niv, Micahel Nossal, Gregory Schnitzer, Jean Stoner, Andrew Kapit and Richard Toren, 2006, “Using Intrinsic and Extrinsic Metrics to Evaluate Accuracy and Facilitation in Computer-assisted Coding”, *Perspectives in Health Information Management Computer Assisted Coding Conference Proceedings*; Fall 2006.
- Roitblat, h. L., Kershaw, A., and Oot, P. “Document categorization in legal electronic discovery: Computer classification vs. manual review”, *Journal of the American Society for Information Science and Technology* 61 (2010), 70–80.
- Özlem Uzuner, PhD,^{a b}*Ira Goldstein, MBA,^a Yuan Luo, MS,^a and Isaac Kohane, MD, PhD^c, “Identifying Patient Smoking Status from Medical Discharge”, *Journal of the American Medical Informatics Association*. 2008 Jan-Feb; 15(1): 14-24.
- Voorhees, Ellen M., “Variations in relevance judgments and the measurement of retrieval effectiveness”, *Information Processing & Management* 36, 5 (2000), 697–716
- Wang, Jianqiang and Dagobert Soergel, “A User Study of Relevance Judgments for E-Discovery”, *ASIST* 2010, October 22-27, 2010, Pittsburgh, PA.

APPENDIX

<table border="1"> <thead> <tr> <th colspan="3">A</th> </tr> <tr> <th></th> <th>R</th> <th>NR</th> <th>Total</th> </tr> </thead> <tbody> <tr> <th>B</th> <td>3698</td> <td>2872</td> <td>6570</td> </tr> <tr> <th>NR</th> <td>153</td> <td>5407</td> <td>5560</td> </tr> <tr> <th>Total</th> <td>3851</td> <td>8279</td> <td>12130</td> </tr> </tbody> </table>			A				R	NR	Total	B	3698	2872	6570	NR	153	5407	5560	Total	3851	8279	12130																																																																																																								
A																																																																																																																													
	R	NR	Total																																																																																																																										
B	3698	2872	6570																																																																																																																										
NR	153	5407	5560																																																																																																																										
Total	3851	8279	12130																																																																																																																										
<table border="1"> <thead> <tr> <th colspan="3">A</th> </tr> <tr> <th></th> <th>R</th> <th>NR</th> <th>Total</th> </tr> </thead> <tbody> <tr> <th>C</th> <td>3142</td> <td>1347</td> <td>4489</td> </tr> <tr> <th>NR</th> <td>709</td> <td>6932</td> <td>7641</td> </tr> <tr> <th>Total</th> <td>3851</td> <td>8279</td> <td>12130</td> </tr> </tbody> </table>		A				R	NR	Total	C	3142	1347	4489	NR	709	6932	7641	Total	3851	8279	12130	<table border="1"> <thead> <tr> <th colspan="3">B</th> </tr> <tr> <th></th> <th>R</th> <th>NR</th> <th>Total</th> </tr> </thead> <tbody> <tr> <th>C</th> <td>4014</td> <td>475</td> <td>4489</td> </tr> <tr> <th>NR</th> <td>2556</td> <td>5085</td> <td>7641</td> </tr> <tr> <th>Total</th> <td>6570</td> <td>5560</td> <td>12130</td> </tr> </tbody> </table>		B				R	NR	Total	C	4014	475	4489	NR	2556	5085	7641	Total	6570	5560	12130																																																																																				
A																																																																																																																													
	R	NR	Total																																																																																																																										
C	3142	1347	4489																																																																																																																										
NR	709	6932	7641																																																																																																																										
Total	3851	8279	12130																																																																																																																										
B																																																																																																																													
	R	NR	Total																																																																																																																										
C	4014	475	4489																																																																																																																										
NR	2556	5085	7641																																																																																																																										
Total	6570	5560	12130																																																																																																																										
<table border="1"> <thead> <tr> <th colspan="3">A</th> </tr> <tr> <th></th> <th>R</th> <th>NR</th> <th>Total</th> </tr> </thead> <tbody> <tr> <th>D</th> <td>1779</td> <td>1020</td> <td>2799</td> </tr> <tr> <th>NR</th> <td>2072</td> <td>7259</td> <td>9331</td> </tr> <tr> <th>Total</th> <td>3851</td> <td>8279</td> <td>12130</td> </tr> </tbody> </table>		A				R	NR	Total	D	1779	1020	2799	NR	2072	7259	9331	Total	3851	8279	12130	<table border="1"> <thead> <tr> <th colspan="3">B</th> </tr> <tr> <th></th> <th>R</th> <th>NR</th> <th>Total</th> </tr> </thead> <tbody> <tr> <th>D</th> <td>2594</td> <td>205</td> <td>2799</td> </tr> <tr> <th>NR</th> <td>3976</td> <td>5355</td> <td>9331</td> </tr> <tr> <th>Total</th> <td>6570</td> <td>5560</td> <td>12130</td> </tr> </tbody> </table>		B				R	NR	Total	D	2594	205	2799	NR	3976	5355	9331	Total	6570	5560	12130	<table border="1"> <thead> <tr> <th colspan="3">C</th> </tr> <tr> <th></th> <th>R</th> <th>NR</th> <th>Total</th> </tr> </thead> <tbody> <tr> <th>D</th> <td>1958</td> <td>841</td> <td>2799</td> </tr> <tr> <th>NR</th> <td>2531</td> <td>6800</td> <td>9331</td> </tr> <tr> <th>Total</th> <td>4489</td> <td>7641</td> <td>12130</td> </tr> </tbody> </table>		C				R	NR	Total	D	1958	841	2799	NR	2531	6800	9331	Total	4489	7641	12130																																																															
A																																																																																																																													
	R	NR	Total																																																																																																																										
D	1779	1020	2799																																																																																																																										
NR	2072	7259	9331																																																																																																																										
Total	3851	8279	12130																																																																																																																										
B																																																																																																																													
	R	NR	Total																																																																																																																										
D	2594	205	2799																																																																																																																										
NR	3976	5355	9331																																																																																																																										
Total	6570	5560	12130																																																																																																																										
C																																																																																																																													
	R	NR	Total																																																																																																																										
D	1958	841	2799																																																																																																																										
NR	2531	6800	9331																																																																																																																										
Total	4489	7641	12130																																																																																																																										
<table border="1"> <thead> <tr> <th colspan="3">A</th> </tr> <tr> <th></th> <th>R</th> <th>NR</th> <th>Total</th> </tr> </thead> <tbody> <tr> <th>E</th> <td>2351</td> <td>937</td> <td>3288</td> </tr> <tr> <th>NR</th> <td>1500</td> <td>7342</td> <td>8842</td> </tr> <tr> <th>Total</th> <td>3851</td> <td>8279</td> <td>12130</td> </tr> </tbody> </table>		A				R	NR	Total	E	2351	937	3288	NR	1500	7342	8842	Total	3851	8279	12130	<table border="1"> <thead> <tr> <th colspan="3">B</th> </tr> <tr> <th></th> <th>R</th> <th>NR</th> <th>Total</th> </tr> </thead> <tbody> <tr> <th>E</th> <td>3228</td> <td>60</td> <td>3288</td> </tr> <tr> <th>NR</th> <td>3342</td> <td>5500</td> <td>8842</td> </tr> <tr> <th>Total</th> <td>6570</td> <td>5560</td> <td>12130</td> </tr> </tbody> </table>		B				R	NR	Total	E	3228	60	3288	NR	3342	5500	8842	Total	6570	5560	12130	<table border="1"> <thead> <tr> <th colspan="3">C</th> </tr> <tr> <th></th> <th>R</th> <th>NR</th> <th>Total</th> </tr> </thead> <tbody> <tr> <th>E</th> <td>2475</td> <td>813</td> <td>3288</td> </tr> <tr> <th>NR</th> <td>2014</td> <td>6828</td> <td>8842</td> </tr> <tr> <th>Total</th> <td>4489</td> <td>7641</td> <td>12130</td> </tr> </tbody> </table>		C				R	NR	Total	E	2475	813	3288	NR	2014	6828	8842	Total	4489	7641	12130	<table border="1"> <thead> <tr> <th colspan="3">D</th> </tr> <tr> <th></th> <th>R</th> <th>NR</th> <th>Total</th> </tr> </thead> <tbody> <tr> <th>E</th> <td>1850</td> <td>1438</td> <td>3288</td> </tr> <tr> <th>NR</th> <td>949</td> <td>7893</td> <td>8842</td> </tr> <tr> <th>Total</th> <td>2799</td> <td>9331</td> <td>12130</td> </tr> </tbody> </table>		D				R	NR	Total	E	1850	1438	3288	NR	949	7893	8842	Total	2799	9331	12130																																										
A																																																																																																																													
	R	NR	Total																																																																																																																										
E	2351	937	3288																																																																																																																										
NR	1500	7342	8842																																																																																																																										
Total	3851	8279	12130																																																																																																																										
B																																																																																																																													
	R	NR	Total																																																																																																																										
E	3228	60	3288																																																																																																																										
NR	3342	5500	8842																																																																																																																										
Total	6570	5560	12130																																																																																																																										
C																																																																																																																													
	R	NR	Total																																																																																																																										
E	2475	813	3288																																																																																																																										
NR	2014	6828	8842																																																																																																																										
Total	4489	7641	12130																																																																																																																										
D																																																																																																																													
	R	NR	Total																																																																																																																										
E	1850	1438	3288																																																																																																																										
NR	949	7893	8842																																																																																																																										
Total	2799	9331	12130																																																																																																																										
<table border="1"> <thead> <tr> <th colspan="3">A</th> </tr> <tr> <th></th> <th>R</th> <th>NR</th> <th>Total</th> </tr> </thead> <tbody> <tr> <th>F</th> <td>3428</td> <td>2648</td> <td>6076</td> </tr> <tr> <th>NR</th> <td>423</td> <td>5631</td> <td>6054</td> </tr> <tr> <th>Total</th> <td>3851</td> <td>8279</td> <td>12130</td> </tr> </tbody> </table>		A				R	NR	Total	F	3428	2648	6076	NR	423	5631	6054	Total	3851	8279	12130	<table border="1"> <thead> <tr> <th colspan="3">B</th> </tr> <tr> <th></th> <th>R</th> <th>NR</th> <th>Total</th> </tr> </thead> <tbody> <tr> <th>F</th> <td>5407</td> <td>669</td> <td>6076</td> </tr> <tr> <th>NR</th> <td>1163</td> <td>4891</td> <td>6054</td> </tr> <tr> <th>Total</th> <td>6570</td> <td>5560</td> <td>12130</td> </tr> </tbody> </table>		B				R	NR	Total	F	5407	669	6076	NR	1163	4891	6054	Total	6570	5560	12130	<table border="1"> <thead> <tr> <th colspan="3">C</th> </tr> <tr> <th></th> <th>R</th> <th>NR</th> <th>Total</th> </tr> </thead> <tbody> <tr> <th>F</th> <td>3779</td> <td>2297</td> <td>6076</td> </tr> <tr> <th>NR</th> <td>710</td> <td>5344</td> <td>6054</td> </tr> <tr> <th>Total</th> <td>4489</td> <td>7641</td> <td>12130</td> </tr> </tbody> </table>		C				R	NR	Total	F	3779	2297	6076	NR	710	5344	6054	Total	4489	7641	12130	<table border="1"> <thead> <tr> <th colspan="3">D</th> </tr> <tr> <th></th> <th>R</th> <th>NR</th> <th>Total</th> </tr> </thead> <tbody> <tr> <th>F</th> <td>2507</td> <td>3569</td> <td>6076</td> </tr> <tr> <th>NR</th> <td>292</td> <td>5762</td> <td>6054</td> </tr> <tr> <th>Total</th> <td>2799</td> <td>9331</td> <td>12130</td> </tr> </tbody> </table>		D				R	NR	Total	F	2507	3569	6076	NR	292	5762	6054	Total	2799	9331	12130	<table border="1"> <thead> <tr> <th colspan="3">E</th> </tr> <tr> <th></th> <th>R</th> <th>NR</th> <th>Total</th> </tr> </thead> <tbody> <tr> <th>F</th> <td>3121</td> <td>2955</td> <td>6076</td> </tr> <tr> <th>NR</th> <td>167</td> <td>5887</td> <td>6054</td> </tr> <tr> <th>Total</th> <td>3288</td> <td>8842</td> <td>12130</td> </tr> </tbody> </table>		E				R	NR	Total	F	3121	2955	6076	NR	167	5887	6054	Total	3288	8842	12130																					
A																																																																																																																													
	R	NR	Total																																																																																																																										
F	3428	2648	6076																																																																																																																										
NR	423	5631	6054																																																																																																																										
Total	3851	8279	12130																																																																																																																										
B																																																																																																																													
	R	NR	Total																																																																																																																										
F	5407	669	6076																																																																																																																										
NR	1163	4891	6054																																																																																																																										
Total	6570	5560	12130																																																																																																																										
C																																																																																																																													
	R	NR	Total																																																																																																																										
F	3779	2297	6076																																																																																																																										
NR	710	5344	6054																																																																																																																										
Total	4489	7641	12130																																																																																																																										
D																																																																																																																													
	R	NR	Total																																																																																																																										
F	2507	3569	6076																																																																																																																										
NR	292	5762	6054																																																																																																																										
Total	2799	9331	12130																																																																																																																										
E																																																																																																																													
	R	NR	Total																																																																																																																										
F	3121	2955	6076																																																																																																																										
NR	167	5887	6054																																																																																																																										
Total	3288	8842	12130																																																																																																																										
<table border="1"> <thead> <tr> <th colspan="3">A</th> </tr> <tr> <th></th> <th>R</th> <th>NR</th> <th>Total</th> </tr> </thead> <tbody> <tr> <th>G</th> <td>2934</td> <td>1880</td> <td>4814</td> </tr> <tr> <th>NR</th> <td>917</td> <td>6399</td> <td>7316</td> </tr> <tr> <th>Total</th> <td>3851</td> <td>8279</td> <td>12130</td> </tr> </tbody> </table>		A				R	NR	Total	G	2934	1880	4814	NR	917	6399	7316	Total	3851	8279	12130	<table border="1"> <thead> <tr> <th colspan="3">B</th> </tr> <tr> <th></th> <th>R</th> <th>NR</th> <th>Total</th> </tr> </thead> <tbody> <tr> <th>G</th> <td>4190</td> <td>624</td> <td>4814</td> </tr> <tr> <th>NR</th> <td>2380</td> <td>4936</td> <td>7316</td> </tr> <tr> <th>Total</th> <td>6570</td> <td>5560</td> <td>12130</td> </tr> </tbody> </table>		B				R	NR	Total	G	4190	624	4814	NR	2380	4936	7316	Total	6570	5560	12130	<table border="1"> <thead> <tr> <th colspan="3">C</th> </tr> <tr> <th></th> <th>R</th> <th>NR</th> <th>Total</th> </tr> </thead> <tbody> <tr> <th>G</th> <td>3081</td> <td>1733</td> <td>4814</td> </tr> <tr> <th>NR</th> <td>1408</td> <td>5908</td> <td>7316</td> </tr> <tr> <th>Total</th> <td>4489</td> <td>7641</td> <td>12130</td> </tr> </tbody> </table>		C				R	NR	Total	G	3081	1733	4814	NR	1408	5908	7316	Total	4489	7641	12130	<table border="1"> <thead> <tr> <th colspan="3">D</th> </tr> <tr> <th></th> <th>R</th> <th>NR</th> <th>Total</th> </tr> </thead> <tbody> <tr> <th>G</th> <td>1829</td> <td>2985</td> <td>4814</td> </tr> <tr> <th>NR</th> <td>970</td> <td>6346</td> <td>7316</td> </tr> <tr> <th>Total</th> <td>2799</td> <td>9331</td> <td>12130</td> </tr> </tbody> </table>		D				R	NR	Total	G	1829	2985	4814	NR	970	6346	7316	Total	2799	9331	12130	<table border="1"> <thead> <tr> <th colspan="3">E</th> </tr> <tr> <th></th> <th>R</th> <th>NR</th> <th>Total</th> </tr> </thead> <tbody> <tr> <th>G</th> <td>2418</td> <td>2396</td> <td>4814</td> </tr> <tr> <th>NR</th> <td>870</td> <td>6446</td> <td>7316</td> </tr> <tr> <th>Total</th> <td>3288</td> <td>8842</td> <td>12130</td> </tr> </tbody> </table>		E				R	NR	Total	G	2418	2396	4814	NR	870	6446	7316	Total	3288	8842	12130	<table border="1"> <thead> <tr> <th colspan="3">F</th> </tr> <tr> <th></th> <th>R</th> <th>NR</th> <th>Total</th> </tr> </thead> <tbody> <tr> <th>G</th> <td>4062</td> <td>752</td> <td>4814</td> </tr> <tr> <th>NR</th> <td>2014</td> <td>5302</td> <td>7316</td> </tr> <tr> <th>Total</th> <td>6076</td> <td>6054</td> <td>12130</td> </tr> </tbody> </table>		F				R	NR	Total	G	4062	752	4814	NR	2014	5302	7316	Total	6076	6054	12130
A																																																																																																																													
	R	NR	Total																																																																																																																										
G	2934	1880	4814																																																																																																																										
NR	917	6399	7316																																																																																																																										
Total	3851	8279	12130																																																																																																																										
B																																																																																																																													
	R	NR	Total																																																																																																																										
G	4190	624	4814																																																																																																																										
NR	2380	4936	7316																																																																																																																										
Total	6570	5560	12130																																																																																																																										
C																																																																																																																													
	R	NR	Total																																																																																																																										
G	3081	1733	4814																																																																																																																										
NR	1408	5908	7316																																																																																																																										
Total	4489	7641	12130																																																																																																																										
D																																																																																																																													
	R	NR	Total																																																																																																																										
G	1829	2985	4814																																																																																																																										
NR	970	6346	7316																																																																																																																										
Total	2799	9331	12130																																																																																																																										
E																																																																																																																													
	R	NR	Total																																																																																																																										
G	2418	2396	4814																																																																																																																										
NR	870	6446	7316																																																																																																																										
Total	3288	8842	12130																																																																																																																										
F																																																																																																																													
	R	NR	Total																																																																																																																										
G	4062	752	4814																																																																																																																										
NR	2014	5302	7316																																																																																																																										
Total	6076	6054	12130																																																																																																																										

TABLE 9 – THE CONTINGENT RELATIONS BETWEEN THE RESPONSIVENESS CODING OF SEVEN REVIEW TEAMS