Louisiana Law Review

Volume 74 | Number 2
Eastern District of Louisiana: The Nation's MDL Laboratory - A Symposium
Winter 2014

Predictive Coding: Taking the Devil Out of the Details

L. Casey Auttonberry

Predictive Coding: Taking the Devil Out of the Details

INTRODUCTION

Discovery has changed, and electronically stored information (ESI) was the catalyst.[1] Though "[e]-discovery matters are no longer the novel issues that they once were,"[2] technology is constantly changing.[3] It was estimated that in 2009 there were 988 exabytes of data in existence, an amount that would stretch from the Sun to Pluto and back in paper form.[4] Massive amounts of ESI have become a huge problem in litigation.[5] Organizations are retaining more information than ever,[6] and lawsuits among these organizations sometimes require lawyers to review more than 100 million documents.[7] Trying to find relevant information in so much

---

1. *See* George L. Paul & Jason R. Baron, *Information Inflation: Can the Legal System Adapt?*, 13 RICH. J.L. & TECH. 10 (2007) (noting that information has changed and that change has affected litigation).

2. Johnson v. Big Lots Stores, Inc., 253 F.R.D. 381, 395 (E.D. La. 2008). Electronic discovery (e-discovery) is "[t]he process of identifying, collecting, processing, analyzing, and reviewing ESI for legal proceedings . . . ." RECOMMIND, INC., PREDICTIVE CODING FOR DUMMIES 3–4 (Recommind Spec. Ed., 2013), *available at* http://media.wiley.com/assets/7072/74/9781118522301_custom.pdf.

3. Technology can also change the way litigation is handled. For instance, the "miracle of photographic reproduction" markedly reduced the burden of transporting discovery materials. The Sedona Conference, *The Sedona Conference Commentary on Proportionality in Electronic Discovery*, 11 SEDONA CONF. J. 289, 292 (2010). The Sedona Conference is a non-profit research and educational organization that seeks "the reasoned and just advancement of law" by creating a "think-tank" setting for leaders of the legal community to discuss current issues in the legal practice. THE SEDONA CONF., https://thesedonaconference.org/aboutus (last visited Oct. 27, 2013).

4. Bennett B. Borden et al., *Four Years Later: How the 2006 Amendments to the Federal Rules Have Reshaped the E-discovery Landscape and Are Revitalizing the Civil Justice System*, 17 RICH. J.L. & TECH. 10, 14 (2011) (citing Jason R. Baron & Ralph C. Losey, *e-Discovery: Did You Know?*, YOUTUBE (Feb. 11, 2010), http://www.youtube.com/watch?v=bWbJWcsPp1M&feature=player_embedded). To put the term "exabyte" in perspective, one exabyte is equal to "a billion billion bytes." DICTIONARY.COM, http://dictionary.reference.com/browse/exabyte?s=t (last visited Oct. 27, 2013).

5. Paul & Baron, *supra* note 1, at 1–2.

6. *Id*. at 1 n.2 (citing GEORGE L. PAUL & BRUCE H. NEARON, THE DISCOVERY REVOLUTION: E-DISCOVERY AMENDMENTS TO THE FEDERAL RULES OF CIVIL PROCEDURE 4–5 (2d ed. 2006) (stating that companies are retaining thousands of times more information now than a few decades ago)).

7. The Sedona Conference, *The Case for Cooperation*, 10 SEDONA CONF. J. 339, 356 (2009) [hereinafter *Case for Cooperation*].

ESI during e-discovery can be grueling for lawyers and expensive for clients.[8]

In the past, lawyers could conduct effective discovery using only manual review.[9] Now, with the increased amount of information retained by parties to litigation, using only manual review in e-discovery is not a realistic option.[10] One way lawyers have dealt with ESI is by using keyword searches, which have become the norm in e-discovery because they allow lawyers to more easily navigate through electronic information.[11] However, even with keyword searches, the amount of ESI can still sometimes be unmanageable.[12]

Fortunately, there are some strategies that litigants can use to make searching through ESI more manageable. One strategy involves new technological tools in e-discovery.[13] One of these tools, called "predictive coding,"[14] could "fundamentally change"

---

8. *See* Paul & Baron, *supra* note 1, at 1–2 (describing how the massive amount of discoverable information has "stressed the legal system" and made litigation "prohibitively expensive").

9. *See* William W. Belt et al., *Technology-Assisted Document Review: Is It Defensible?*, 18 RICH. J.L. & TECH. 10, 2 (2012) (discussing how discovery materials are now sent on hard drives instead of in boxes). Manual review requires humans to read through documents one at a time and classify them as relevant or irrelevant to the document request. Maura R. Grossman & Gordon V. Cormack, *Technology-Assisted Review in E-Discovery Can Be More Effective and More Efficient Than Exhaustive Manual Review*, 17 RICH. J.L. & TECH. 11, 2 (2011).

10. *See* Andrew Peck, *Search, Forward: Will Manual Document Review and Keyword Searches be Replaced by Computer-Assisted Coding?*, L. TECH. NEWS (Oct. 2011), http://www.recommind.com/sites/default/files/LTN_Search_Forward_Peck_Recommind.pdf ("[T]he volume of electronically stored information . . . has largely eliminated manual review as the sole method of document review . . . ."); *see also* Paul & Baron, *supra* note 1, at 3 (noting that "[l]itigators can no longer depend on manual review alone").

11. The Sedona Conference, *The Sedona Conference Best Practices Commentary on the Use of Search and Information Retrieval Methods in E-Discovery*, 8 SEDONA CONF. J. 189, 200 (2007) [hereinafter *Sedona Conference Best Practices*]. In a keyword search, the human searcher inputs "words into a computer which then retrieves documents within the collection containing the same words." This method is also known as "Boolean searching." MATTHEW D. NELSON, ESQ., PREDICTIVE CODING FOR DUMMIES 9 (Symantic Spec. Ed., 2012), *available at* http://media.wiley.com/assets/7056/00/9781118482377_custom.pdf. Keyword searches allow more advanced searches using multiple word combinations and root word derivatives. *Id.*

12. *See* Jason R. Baron, *Law in the Age of Exabytes: Some Further Thoughts on 'Information Inflation' and Current Issues in E-Discovery Search*, 17 RICH. J.L. & TECH. 9, 10 (2011).

13. *See* Paul & Baron, *supra* note 1, at 26.

14. Melissa Whittingham et al., *Predictive Coding: E-Discovery Game Changer?*, EDDE J. 11 (2011), *available at* http://www.cov.com/files/Publication/9f38beae-2753-481d-b638-55f86c46931f/Presentation/PublicationAttachment/

discovery in litigation involving large amounts of ESI.[15] Predictive coding is a "machine-learning technology" that, with a relatively small amount of human input, teaches a computer to "predict" document classification.[16] The coding tool uses a man-made "definition" to make "rules" for classifying documents[17] and then organizes the documents within a larger document collection based on how well they match the man-made definition and rules.[18] The end result is that lawyers manually review a much smaller set of documents.[19] Predictive coding therefore effectively "alleviat[es] the need to review whole masses of records in order to find the relevant few."[20] Most importantly, predictive coding is estimated to reduce e-discovery costs as much as 45% to 71% while maintaining search quality.[21] Studies suggest that technology-assisted review is no less accurate than human review.[22]

6e933c53-08a3-4f05-8962-587348107592/Predictive%20Coding%20-%20E-Discovery%20Game%20Changer.pdf. Predictive coding is also known as "automated document review, automated document classification, automatic categorization, predictive categorization, and predictive ranking." *Id*.

15. Scott Vernick, *Predictive Coding: Three Things You Need to Know About This Year's Biggest Legal Tech Trend*, HUFF POST TECH. BLOG (Aug. 15, 2012, 6:36 PM), http://www.huffingtonpost.com/scott-vernick/three-things-you-need-to-_b_1773959.html.

16. NELSON, *supra* note 11, at 7. Predictive coding can also be described as "technology-assisted review," which is a search process in which humans use technology to find responsive documents in a large data collection. Grossman & Cormack, *supra* note 9, at 2.

17. Chuck Rothman, *What is this Predictive Coding Thing Anyway?*, EDISCOVERYJOURNAL.COM (Mar. 14, 2012, 8:00 AM), http://ediscoveryjournal.com/2012/03/what-is-this-predictive-coding-thing-anyway/. These "definitions" are called "classifiers." *Id.* Humans review a small set of documents and determine their relevance to the case's facts to formulate the definition for the predictive coding tool. Ari Kaplan & Joe Looby, *Advice from Counsel: Can Predictive Coding Deliver on Its Promise?*, FTI CONSULTING TECHN. 1 (2012), *available at* http://www.ftitechnology.com/doc/White-Papers/whitepaper-2012-Predictive-Coding-Survey.pdf [hereinafter *Advice from Counsel*]. The person actually conducting the coding process may vary depending on the situation. *See infra* Part II.

18. Rothman, *supra* note 17. Several other steps are necessary for the predictive coding tool to find documents effectively. For example, the searcher uses an "iterative approach" in the process, which incorporates "document sampling and quality assurance" checks. *Advice from Counsel*, *supra* note 17, at 1. These steps are discussed further *infra* Parts II–IV.

19. *See* Grossman & Cormack, *supra* note 9, at 2.

20. Rothman, *supra* note 17.

21. EDISCOVERY INSTITUTE SURVEY ON PREDICTIVE CODING 3 (2010), *available at* http://www.discovia.com/wp-content/uploads/2012/07/2010_EDI_PredictiveCodingSurvey.pdf [hereinafter *EDI Survey*].

22. *See, e.g.*, Herbert L. Roitblat et al., *Document Categorization in Legal Electronic Discovery: Computer Classification vs. Manual Review*, J. AM. SOC'Y

Although predictive coding is relatively new in the legal realm, litigants are quickly realizing its utility as an e-discovery tool. With more use of the technology, courts are now endorsing predictive coding protocols, thereby legitimizing its use in litigation.[23] However, judicially approved predictive coding protocols have employed the technology very differently.[24] Implementing an effective predictive coding protocol involves many considerations,[25] and because e-discovery tools "are only effective if used properly,"[26] the specifics of predictive coding protocols leave much room for dispute.[27]

Another strategy that can be used to make searching through ESI more manageable is a cooperative approach to e-discovery, which emphasizes cooperation, transparency, and efficiency.[28] The cooperative approach to e-discovery has been very effective at

---

INFO. SCI. & TECH. 70 (2010), *available at* http://www.ediscoveryinstitute.org/images/uploaded/592.pdf. The results of this study "support[ed] the idea that machine categorization is no less accurate at identifying relevant/responsive documents than employing a team of reviewers." *Id.* Additionally, the idea that human manual review is the "gold standard" by which search accuracy should be measured is increasingly regarded as more mythical than factual. *See Sedona Conference Best Practices*, *supra* note 11, at 199 (discussing how research suggests that manual review is not necessarily the "gold standard" that it has traditionally been considered).

23. *See, e.g.*, Moore v. Publicis Groupe, 287 F.R.D. 182 (S.D.N.Y. 2012); *In re* Actos (Pioglitazone) Prods. Liab. Litig., No. 6:11-md-2299, 2012 WL 6061973 (W.D. La. July 27, 2012) (Case Management Order: Protocol Relating to the Production of Electronically Stored Information). A protocol is also known as a workflow. *See, e.g.*, RECOMMIND, INC., *supra* note 2, at 18; *see also* discussion *infra* note 106 and accompanying text.

24. *See infra* Parts II–IV.

25. *See* NELSON, *supra* note 11, at 31 (noting that those using predictive coding must make complex decisions when designing a workflow).

26. *Id.* at 42.

27. Jan Conlin & Andrew Pieper, *Litigation: Predictive Coding's Grand Debut*, INSIDECOUNSEL (Sept. 13, 2012), http://www.insidecounsel.com/2012/09/13/litigation-predictive-codings-grand-debut. Areas that may cause disputes include the use of the technology in general as well as the specifics of operating the coding software as detailed in protocols. *Id.*

28. *See* Paul & Baron, *supra* note 1, at 26. Cooperation has been emphasized in e-discovery efforts that accompany new e-discovery challenges. *See generally, e.g.*, The Sedona Conference, *The Sedona Conference Cooperation Proclamation*, 10 SEDONA CONF. J. 331 (2009) [hereinafter *Cooperation Proclamation*]. Transparency among counsel is a critical part of the cooperative effort. *Moore*, 287 F.R.D. at 191. Additionally, the promise of predictive coding is to make e-discovery more efficient by reducing discovery costs. Whittingham, *supra* note 14.

addressing disputes[29] and should be at the heart of every e-discovery effort.[30] Therefore, this Comment argues that parties and courts should implement predictive coding protocols that are consistent with the cooperative approach to e-discovery.[31] By considering the predictive coding protocols in *Moore v. Publicis Groupe*[32] and *In re Actos (Pioglitazone) Products Liability Litigation*,[33] this Comment offers a model protocol to effectuate the most cooperative, transparent, and efficient methods of conducting predictive coding in e-discovery.

Part I considers the importance of cooperation, transparency, and efficiency in e-discovery and how these principles can affect predictive coding. Part II summarizes the predictive coding protocols established in *Moore v. Publicis Groupe*[34] and *In re Actos (Pioglitazone) Products Liability Litigation*.[35] Part III points out the advantages and deficiencies of the protocols in conjunction with the discovery principles of cooperation, transparency, and efficiency. Finally, by focusing on those principles, Part IV proposes a model protocol to reduce unnecessary disputes when parties decide to use predictive coding.

I. DEALING WITH THE ESI ENIGMA: THE COOPERATIVE APPROACH

Due to technological advances, the volume of discoverable information has drastically increased.[36] This "information inflation" has made searching for relevant information incredibly expensive and challenging.[37] New approaches are necessary to deal with these 21st century e-discovery problems.[38] One of these approaches,

---

29. Brian C. Vick & Neil C. Magnuson, *The Promise of a Cooperative and Proportional Discovery Process in North Carolina: House Bill 380 and the New State Electronic Discovery Rules*, 34 CAMPBELL L. REV. 233, 249 (2012).

30. Paul & Baron, *supra* note 1, at 25.

31. Generally, courts will allow parties to work according to a discovery agreement as long as the agreement is reasonable and explainable. *Sedona Conference Best Practices*, *supra* note 11, at 204. However, while courts liberally honor discovery agreements and protocols, "[t]he desirability of some judicial control of discovery can hardly be doubted." FED. R. CIV. P. 26(f) advisory committee's note.

32. 287 F.R.D. 182.

33. *In re* Actos (Pioglitazone) Prods. Liab. Litig., No. 6:11-md-2299, 2012 WL 6061973 (W.D. La. July 27, 2012) (Case Management Order: Protocol Relating to the Production of Electronically Stored Information).

34. 287 F.R.D. 182.

35. *In re Actos*, 2012 WL 6061973.

36. Paul & Baron, *supra* note 1, at 11–13.

37. *Id.* at 1.

38. *Id.* at 24.

which this Part discusses, suggests increased cooperation, transparency, and efficiency during e-discovery.[39]

A. Cooperative E-Discovery

The information inflation has caused an urgent need for cooperative e-discovery.[40] In order to cooperate, the legal community must first understand what cooperation means. Like the concepts of "good faith" and the "reasonable man," "cooperation" in the discovery context does not have a precise definition.[41] The most basic concept of cooperation is "a certain level of candor and transparency in communications between counsel."[42] The "certain level" aspect of this concept is important. Cooperation does not aim to eradicate all disputes from the adversarial process, but it does seek to avoid unnecessary disputes.[43] In the past, lawyers inherently understood this type of cooperative effort, but certain aspects of litigation have decreased the consensus understanding of cooperation.[44]

For example, the American legal system's adversarial nature can hinder cooperative discovery.[45] Within the adversarial process is the duty to zealously advocate for clients.[46] Lawyers have a duty to

39. *Id.* at 25.
40. *Case for Cooperation*, *supra* note 7, at 342.
41. *Id.* at 340.
42. *Id.*
43. Ralph C. Losey, *Lawyers Behaving Badly: Understanding Unprofessional Conduct in E-Discovery*, 60 MERCER L. REV. 983, 997 (2009) [hereinafter *Lawyers Behaving Badly*]. *See also Case for Cooperation*, *supra* note 7, at 344 (explaining that legitimate discovery disputes should continue to be pursued, but courts criticize parties who bring unnecessary disputes that could have been avoided by cooperating with opposing parties).
44. *Cooperation Proclamation*, *supra* note 28, at 332 (acknowledging that cooperation and collaboration were "understood" when the Federal Rules of Civil Procedure were adopted in 1938).
45. *See* Wayne Brazil, *The Adversary Character of Civil Discovery: A Critique and Proposals for Change*, 31 VAND. L. REV. 1295, 1296 (1978) ("[T]he adversary character of civil discovery, with substantial reinforcement from the economic structure of our legal system, promotes practices that systematically impede the attainment of the principal purposes for which discovery was designed."). Discovery is not designed to adjudicate the case—It is designed to find the facts surrounding the case. RALPH C. LOSEY, *Tall Tales and Ethics with Karl Schieneman*, *in* ADVENTURES IN ELECTRONIC DISCOVERY 236, 251 (2011 ed.).
46. The idea of "zealous advocacy" derives from the Model Rules of Professional Conduct. Comment 1 to Rule 1.3 describes a lawyer's representative diligence as "zeal in advocacy." MODEL RULES OF PROF'L CONDUCT R. 1.3 cmt. 1 (2012).

advocate for their clients,[47] but many lawyers interpret the term "zealous advocacy" as a "hide the ball mentality."[48] Lawyers justify this mentality with privilege or confidentiality considerations.[49] These considerations, while sometimes legitimate, are too often just an excuse for misconduct.[50]

However, sometimes clients, rather than their attorneys, are responsible for uncooperative e-discovery.[51] When an aggressive client has greater resources than its adversary, e-discovery can be used as a weapon to force weaker parties into unfavorable settlements.[52] This often puts the aggressive client's lawyer in a situation where he or she must choose between conducting cooperative discovery and pleasing the client by conducting discovery uncooperatively.[53] If the lawyer chooses to be uncooperative, opposing counsel often reciprocates, forcing the parties to conduct discovery "the hard way."[54] Moreover, parties who initiate this uncooperative behavior can frustrate others who fully intend to be cooperative, making those with good intentions wonder if cooperation is even worth the effort.[55]

But cooperation *is* worth the effort. Some parties may perceive being uncooperative as a litigation strategy, but they are really only hurting their cause.[56] Gamesmanship and "hiding the ball" waste resources on unnecessary discovery disputes.[57] In some cases, combative behavior unnecessarily lengthens the discovery process and makes costs case determinative.[58] In other cases, the e-discovery costs can actually rise above the amount in controversy.[59]

---

47. The Sedona Conference, *The Sedona Conference Cooperation Guidance for Litigators & In-House Counsel* 2 (2011) (Richard G. Braman ed., et al.), *available at* https://www.google.com/url?sa=t&rct=j&q=&esrc=s&source=web&cd=2&ved=0CDEQFjAB&url=https%3A%2F%2Fthesedonaconference.org%2Fsystem%2Ffiles%2Fsites%2Fsedona.civicactions.net%2Ffiles%2Fprivate%2Fdrupal%2Ffilesys%2Fpublications%2FCooperation_Guidance_for_Litigators_and_In_House_Counsel.pdf [hereinafter *Guidance*].

48. *Cooperation Proclamation*, *supra* note 28, at 332.

49. *Lawyers Behaving Badly*, *supra* note 43, at 997.

50. *Id.* at 991.

51. *See Case for Cooperation*, *supra* note 7, at 359–61.

52. Margaret Rowell Good, *Loyalty to the Process: Advocacy and Ethics in the Age of E-Discovery*, 86 FLA. BAR J. 96, 99 (2012).

53. *Case for Cooperation*, *supra* note 7, at 359.

54. *Id.* at 360.

55. *Cooperation Proclamation*, *supra* note 28, at 332.

56. David Degnan, *Accounting for the Costs of Electronic Discovery*, 12 MINN. J.L. SCI. & TECH. 151, 189 (2011).

57. *Cooperation Proclamation*, *supra* note 28, at 331.

58. *Case for Cooperation*, *supra* note 7, at 356.

59. Borden et al., *supra* note 4, at 17.

Although "zealous advocacy" has become a popular term, lawyers are actually expected to represent the client with "diligence."[60] Clients may want to use discovery as a weapon, but lawyers are not obligated to pursue every possible advantage for their clients or use aggressive tactics in discovery.[61] Diligent representation is not meant to prohibit cooperation,[62] and cooperative discovery does not have to compromise a client's interests.[63] Pragmatically, the party with greater resources is often subject to greater ESI production.[64] If that party acts uncooperatively and opposing counsel reciprocates, it can actually increase the discovery expenses of the party with greater resources instead of giving it a tactical advantage.[65] Indeed, parties acting uncooperatively in e-discovery can lead to "mutually assured destruction."[66]

Cooperative discovery can also help parties maintain goodwill with courts.[67] Courts expect parties to cooperate in e-discovery,[68]

---

60. MODEL RULES OF PROF'L CONDUCT R. 1.3 (2012). The phrase "zeal in advocacy" derives from the comments to the Model Rules of Professional Conduct. *See supra* note 46 and accompanying text. Some commentators suggest that this phrase actually encourages aggressive, "Rambo" style litigation tactics instead of cooperative and professional behavior. Allen K. Harris, *Increasing Ethics, Professionalism and Civility: Key to Preserving the American Common Law and Adversarial Systems*, 2005 PROF. LAW. 91, 108 (2005).

61. *See* MODEL RULES OF PROF'L CONDUCT R. 1.3 cmt. 1 (2012). Lawyers do not have "to press for every advantage that might be realized for a client," and "diligence does not require the use of offensive tactics . . . ." *Id.* In fact, lawyers should "exercise professional discretion" when representing clients during e-discovery and other aspects of litigation. *Id.*

62. *Lawyers Behaving Badly*, *supra* note 43, at 997.

63. *Case for Cooperation*, *supra* note 7, at 344.

64. Good, *supra* note 52, at 99.

65. *Id.* Cooperation actually helps parties reduce costs by allowing them to get to the merits more quickly and maintain greater control of the case. *Case for Cooperation*, *supra* note 7, at 339. Uncooperative behavior has greater consequences than just increased discovery costs; this behavior in e-discovery has caused an increase in motions for sanctions. RALPH C. LOSEY, *What is Wrong, or Right, with e-Discovery in America*, *in* ADVENTURES IN ELECTRONIC DISCOVERY, *supra* note 45, at 16 (2011 ed.). In 2009, courts heard more e-discovery sanction cases "than in all years prior to 2005 combined." Dan H. Willoughby, Jr. et al., *Sanctions for E-Discovery Violations: By the Numbers*, 60 DUKE L.J. 789, 794 (2010). Unfortunately, even sanction motions are often used as a litigation tactic to increase discovery costs. LOSEY, *supra*, at 16.

66. Borden et al., *supra* note 4, at 17.

67. *Case for Cooperation*, *supra* note 7, at 339.

68. *See* Gareth Evans et al., *2012 Year-End Electronic Discovery and Information Law Update*, LEXOLOGY (Jan. 14, 2013), http://www.lexology.com/library/detail.aspx?g=22518c74-58c9-4551-9959-3e7e289f7fb2 (describing how courts deliberately focused on making e-discovery more cooperative in 2012).

and many courts use *The Sedona Conference Cooperation Proclamation* (the "*Cooperation Proclamation*")[69] as the model standard for cooperative discovery.[70] Jurists have also argued for cooperation in the predictive coding process. In his article *Search, Forward: Will Manual Document Review and Keyword Searches be Replaced by Computer-Assisted Coding?*, U.S. Magistrate Judge Andrew Peck argues that the best way for counsel to engage in the coding process is to follow the model of cooperation set forth in the *Cooperation Proclamation*.[71] Cooperative e-discovery is indeed an integral step toward making the courts "a place where justice may be reached by all,"[72] but it is not the only step.

B. Transparent E-Discovery

The *Cooperation Proclamation* aspires to "facilitate cooperative, collaborative, [and] *transparent* discovery."[73] Transparency is defined as "openness" or "clarity."[74] In the e-discovery context, this means that producing parties should give the court and opposing parties clear and comprehensive explanations of its search processes.[75] Transparency is a vital part of the cooperative effort,[76] and "[a]ll cooperative efforts, actually, should be transparent."[77] Like cooperation, transparent discovery is sometimes hindered by the notion of zealous advocacy and claims of privilege, but there are

---

69. *Cooperation Proclamation*, *supra* note 28.

70. *See, e.g.*, S.E.C. v. Collins & Aikman Corp., 256 F.R.D. 403, 415 (S.D.N.Y. 2009) (directing the parties' attentions to the *Cooperation Proclamation*'s call for cooperative discovery); DeGeer v. Gillis, 755 F. Supp. 2d 909, 918 (N.D. Ill. 2010) (noting that the court endorses the "cooperative, collaborative, and transparent discovery" that the *Cooperation Proclamation* encourages (quoting *Cooperation Proclamation*, *supra* note 28, at 331)).

71. Peck, *supra* note 10. Additionally, the *Cooperation Proclamation* itself encourages parties to cooperate when approaching e-discovery. *See Cooperation Proclamation*, *supra* note 28, at 332 (suggesting that parties work together when "developing automated search and retrieval methodologies to cull relevant information").

72. Degnan, *supra* note 56, at 189.

73. *Cooperation Proclamation*, *supra* note 28, at 331 (emphasis added).

74. BLACK'S LAW DICTIONARY 1638 (9th ed. 2009).

75. The Sedona Conference, *The Sedona Conference Commentary on Achieving Quality in the E-Discovery Process*, 10 SEDONA CONF. J. 299, 307 (2009) [hereinafter *Achieving Quality*].

76. *See* Moore v. Publicis Groupe, 287 F.R.D. 182, 192 (S.D.N.Y. 2012) (discussing why transparency was such an important aspect in the court's approval of the predictive coding protocol).

77. *Guidance*, *supra* note 47, at 1.

several reasons why transparency is necessary now more than ever.[78]

First, the sheer amount of ESI involved in complex cases makes it difficult to "preserve, search, review, and produce" information as necessary for e-discovery.[79] It is imperative that parties act transparently to keep discovery within the bounds of the Federal Rules of Civil Procedure (FRCP).[80] Second, with new technology like predictive coding becoming more prominent in e-discovery, transparency by the technology's proponent can help the court and other parties determine if the technology is a "reasonable" way to produce documents.[81]

Transparency should not end at the court's approval of the search method, though. It begins at the parties' discovery conference,[82] continues through court approval, and lasts the entire duration of the document retrieval, review, and production processes.[83] In effect, transparency throughout the process can enhance the effectiveness of the search by increasing the amount of responsive documents found and reducing the amount of non-responsive documents reviewed.[84]

Third, since tools like predictive coding often require lawyers to consult with software experts more than with opposing counsel, locating and producing ESI is often naturally less transparent than

---

78. *See* Craig B. Shaffer, *"Defensible" By What Standard?*, 13 SEDONA CONF. J. 217, 224 (2012); *Lawyers Behaving Badly*, *supra* note 43, at 991 (discussing the false belief that transparency hinders the ability to zealously advocate for clients). One of the most important reasons that transparency is currently necessary is that the proponents of e-discovery protocols may be required to defend their techniques. *See generally* Shaffer, *supra*.

79. *Case for Cooperation*, *supra* note 7, at 340.

80. Paul & Baron, *supra* note 1, at 35. For example, FRCP 26(b)(2)(C)(iii) recognizes the burdens of e-discovery and requires a court to determine the proportionality of discovery by "considering the needs of the case, the amount in controversy, the parties' resources, the importance of the issues at stake in the action, and the importance of the discovery in resolving the issues." FED. R. CIV. P. 26(b)(2)(C)(iii).

81. *Moore*, 287 F.R.D. 182, 192 (Defendant's "transparency in its proposed ESI search protocol made it easier for the [c]ourt to approve the use of predictive coding."). The term "reasonable," when used in discovery, means that the search method does an adequate job of identifying responsive documents without undue burden. The idea of reasonableness comes from the FRCP. *See* FED. R. CIV. P. 26(g)(1)(B)(iii) (stating that discovery requests, responses, and objections should be "neither unreasonable nor unduly burdensome or expensive"). Discovery does not aim to be perfect, but it should be reasonable. Roitblat et al., *supra* note 22, at 72.

82. *See* FED. R. CIV. P. 26(f)(1) (requiring the parties to "confer as soon as practicable").

83. Paul & Baron, *supra* note 1, at 56 n.134.

84. *Id.*

production of traditional paper materials.[85] Finally, transparency can help reduce motion practice.[86] Transparency replaces "gamesmanship" in e-discovery and allows parties to understand the adversary's reasoning, enabling e-discovery disputes to be resolved more easily.[87] Therefore, transparency gives parties less incentive to compel further discovery through the court[88] and is largely viewed as the preferred manner to reduce e-discovery challenges.[89]

C. Efficient E-Discovery

E-discovery commentators have proposed—and common sense advocates—that modern discovery should aim to find the information relevant to the case "as quickly and efficiently as possible."[90] Due to an increasing volume of ESI, the e-discovery review process has evolved.[91] Manual review alone is no longer an efficient way to conduct a search for relevant documents.[92] One way parties can increase the efficiency of e-discovery is through technology-assisted review.[93] The ultimate goal of these searches is to "produce high recall and high precision (in a cost-effective way)."[94]

---

85. *See* Richard Esenberg, *A Modest Proposal for Human Limitations on Cyberdiscovery*, 64 FLA. L. REV. 965, 970 (2012). E-discovery can become extremely complicated, often resulting in equally complex disputes about what documents are readily obtainable, leading to "discovery about discovery." *Id.* (quoting Paul W. Grimm et al., *Discovery About Discovery: Does the Attorney-Client Privilege Protect All Attorney-Client Communications Relating to the Preservation of Potentially Relevant Information?*, 37 U. BALT. L. REV. 413, 426 (2008)).

86. *See Guidance*, *supra* note 47, at 26–27.

87. *Case for Cooperation*, *supra* note 7, at 344–45.

88. *Guidance*, *supra* note 47, at 26–27.

89. *Sedona Conference Best Practices*, *supra* note 11, at 204.

90. Ralph C. Losey, *Child's Game of 'Go Fish' Is a Poor Model for E-Discovery Search*, E-DISCOVERY TEAM BLOG (Oct. 4, 2009, 4:09 PM), http://e-discoveryteam.com/2009/10/04/childs-game-of-go-fish-is-a-poor-model-for-e-discovery-search/ [hereinafter *Go Fish*]. While this proposal is insightful, it is essentially just a restatement of FRCP 1, which demands the "speedy" and "inexpensive determination of every action." FED. R. CIV. P. 1.

91. *Sedona Conference Best Practices*, *supra* note 11, at 193.

92. *See supra* note 10 and accompanying text.

93. KPMG ENTERPRISE-LEVEL ELECTRONIC DISCOVERY 6 (2012), *available at* http://www.kpmg.com/US/en/IssuesAndInsights/ArticlesPublications/Documents/enterprise-level-overview.pdf.

94. Peck, *supra* note 10. "Recall is the fraction of relevant documents identified during a review, i.e., a measure of completeness. Precision is the fraction of identified documents that are relevant, i.e., it is a measure of accuracy or correctness." *Id.*

In response to the growing amount of ESI, keyword searches became commonly used to increase efficiency in e-discovery.[95] While keyword searching can be a valuable asset, it certainly has its deficiencies,[96] and parties' increasing struggles with the "time and cost requirements" of electronic searches—even with keyword searches—negatively impact efficiency in e-discovery.[97] The newest stage in the evolution of electronic search is predictive coding, which can be a powerful tool to make searching through ESI more manageable and efficient.[98] Many lawyers have wasted no time turning to this new technology to decrease costs and increase efficiency.[99] Even judges have expressed their approval of predictive coding to increase e-discovery efficiency, with the "speedy" and "inexpensive" goals of FRCP 1 as justification for the technology's use.[100]

Another way parties can increase efficiency in e-discovery, regardless of what technology is being used, is through the use of search protocols, also called "workflows."[101] Well-designed protocols can effectively decrease some of the costs and delays associated with e-discovery.[102] When parties are in agreement with a search process and reasonably explain it, courts will usually endorse the protocol and let the parties work according to their agreement.[103] However, protocols can be complex and confusing, particularly when implemented with new technology like predictive coding.[104]

The next Part of this Comment provides two examples of protocols that incorporate predictive coding. It is true that both predictive coding and search protocols can help increase efficiency in e-discovery. However, when two judicially endorsed search protocols using essentially the same technology differ so greatly, it

---

95. *Sedona Conference Best Practices*, *supra* note 11, at 200.
96. *See id.* at 194.
97. Belt, *supra* note 9, at 1.
98. *Id.* Predictive coding is a "scientific analysis that is accompanied by a methodology," while keyword searches are essentially a "bold guess." Jan Puzicha, *Predictive Coding Explained*, RECOMMIND, INC. 14 (2012) (quoting Judge Paul Grimm), *available at* http://www.recommind.com/resources/knowledge_library/predictive-coding-explained-dr-jan-puzicha-0. Removing the guessing aspect from the search process reduces the amount of documents subject to manual review and eliminates the cost of paying lawyers to conduct that manual review. NELSON, *supra* note 11, at 8.
99. Peck, *supra* note 10.
100. *Id.*; FED. R. CIV. P. 1.
101. FED. R. CIV. P. 26(f) advisory committee's note.
102. *Id.*
103. *Sedona Conference Best Practices*, *supra* note 11, at 204.
104. NELSON, *supra* note 11, at 31.

is questionable whether both adequately embrace the cooperative approach to e-discovery.

II. TWO JUDICIALLY APPROVED PREDICTIVE CODING PROTOCOLS

Developing an effective predictive coding protocol can be challenging.[105] Parties usually choose between two general methods of employing predictive coding: the assisted-review method and the comparison method.[106] In the assisted-review method, the user first creates a "seed set," which is a targeted document collection developed using keyword searches.[107] The coders then further develop the seed set in a series of smaller searches called "iterative training," or "iterative rounds."[108] The coders test those results for accuracy, then conduct a full-scale final search before producing the responsive documents to the opposing party.[109] The comparison method is similar to the assisted-review method, except the user creates a "control set," which is a random document collection, instead of a targeted seed set.[110] This simple difference can have serious consequences. Additionally, each step in the protocol can be executed differently. A few of these differences are observed in this Part, which describes the predictive coding protocols endorsed in *Moore v. Publicis Groupe & MSL*[111] and *In re Actos (Pioglitazone) Products Liability Litigation*.[112]

---

105. *See id.* at 31 ("[I]mplementing a proper workflow can be confusing and complex."). Some of this difficulty arises because predictive coding providers can offer varying options when it comes time to implement the workflow. *Id.* at 14. More likely, though, the difficulty is attributable to the newness of predictive coding in the legal field. *Id.*

106. RECOMMIND, INC., *supra* note 2, at 19.

107. *Id.* During the creation of the seed set, both responsive and non-responsive documents are placed in the seed set to teach the coding software the difference between them. *See* NELSON, *supra* note 11, at 17–18.

108. RECOMMIND, INC., *supra* note 2, at 19.

109. *Id.*

110. *Id.* The control set is used as a baseline measurement for measuring the predictive coding tool's training progress. NELSON, *supra* note 11, at 13. The tool's training results are measured against the control set to evaluate the coding tool's performance. *Id.*

111. 287 F.R.D. 182 (S.D.N.Y. 2012).

112. *In re* Actos (Pioglitazone) Prods. Liab. Litig., No. 6:11-md-2299, 2012 WL 6061973 (W.D. La. July 27, 2012) (Case Management Order: Protocol Relating to the Production of Electronically Stored Information).

A. Moore v. Publicis Groupe & MSL

1. Case Background

In *Moore v. Publicis Groupe & MSL*,[113] five plaintiffs sued their employer, Publicis Groupe, and its subsidiary, MSL, for "systemic, company-wide gender discrimination" against its female employees.[114] The court approved an e-discovery protocol as part of MSL's response to the plaintiffs' document requests,[115] which necessitated a review of more than three million e-mails by defendants' counsel.[116] Instead of using more traditional e-discovery search techniques, MSL chose a private third-party predictive coding vendor and implemented the assisted-review method.[117] Because defense counsel elected to utilize predictive coding, plaintiffs requested "clarification" on the way that MSL planned to use the technology.[118]

2. The Predictive Coding Protocol

To begin the predictive coding process, MSL created a seed set.[119] MSL used the predictive coding tool to create an initial 2,399-document "random sample" from the collection of discoverable e-mails.[120] MSL's lawyers reviewed and gave this sample to opposing counsel for review and further input.[121] MSL then created and refined the seed set using keyword searches.[122] MSL reviewed and

113. *Moore*, 287 F.R.D. 182.
114. *Id.* at 183.
115. *Id.* at 187.
116. *Id.* at 184. Additionally, MSL initially proposed that the production of documents be limited to the top 40,000 documents. *Id.* at 185. The court denied this proposal on considerations of proportionality, noting that "if stopping at 40,000 [documents] is going to leave a tremendous number of likely highly responsive documents unproduced, MSL's proposed cutoff doesn't work." *Id.*
117. *Id.* at 199. MSL employed Recommind, who uses a coding tool called Axcelerate. *Id.* Different vendors' products can provide different capabilities to predictive coding consumers, but this Comment does not endorse any specific predictive coding vendor or software. For more information about Recommind's Axcelerate software, see RECOMMIND, http://www.recommind.com/ (last visited Oct. 27, 2013).
118. *Moore*, 287 F.R.D. at 185 (noting that plaintiffs did not object to the use of the technology but, rather, were concerned with how it would be used in that case).
119. *Id.* at 200. This seed set was used to "train" the software to find relevant documents. *Id.* at 200–01.
120. *Id.* at 186.
121. *Id.* at 201.
122. *Id.*

coded the documents resulting from the keyword search and gave those documents to plaintiffs for review. Plaintiffs could give feedback on any documents they thought were coded incorrectly and were to promptly return the documents to MSL.[123] Plaintiffs also provided a supplemental list of keywords for the defendants to search.[124] MSL repeated this process using plaintiffs' supplemental keywords, reviewed and coded 4,000 random documents resulting from the keyword search, and again provided the documents to the plaintiffs to review and return.[125]

Next, MSL started the iterative rounds of training.[126] The predictive coding tool used the seed sets to find similar documents in the collection of e-mails.[127] MSL's lawyers reviewed and coded a 500-document sample that the coding software suggested matched the documents in the seed set.[128] The purpose of this step is to ensure that the software is operating correctly and to adjust it or make changes if necessary.[129] MSL gave the plaintiffs both the

123. *Id.*
124. *Id.*
125. *Id.*
126. *Id.* at 201–02. The purpose of the iterative or successive rounds of training is to stabilize the coding tool and prepare it for an accurate final search. *Id.* at 187. As part of the training process, and all other aspects of the protocol, the parties entered into confidentiality stipulations and clawback agreements. *Id.* at 194. A clawback agreement is a non-waiver agreement in which adversarial parties stipulate that privileged documents inadvertently produced during discovery will be returned to the producing party and do not constitute waiver of privilege. Jessica Wang, *Nonwaiver Agreements after Federal Rule of Evidence 502: A Glance at Quick-Peek and Clawback Agreements*, 56 UCLA L. REV. 1835, 1842 (2009). These agreements are necessary in e-discovery because the amount of ESI subject to production makes inadvertent production of a privileged document virtually inevitable. Matthew A. Reiber, *Latching onto Laches: A Rules-Based Alternative for Resolving Questions of Waiver Following the Inadvertent Production of Privileged Documents in Federal Court Actions*, 38 N.M. L. REV. 197, 198 (2008) (commenting on the inadvertent disclosures of privileged documents during discovery). Clawback agreements are impliedly authorized in FRCP 26(b)(5) and acknowledged with approval in the 2006 Advisory Committee Notes. *See* FED. R. CIV. P. 26(b)(5); FED. R. CIV. P. 26(b)(5) advisory committee's note. Additionally, Federal Rule of Evidence 502(d) acknowledges that these agreements are enforceable and do not constitute waiver of privilege between the parties, and the agreements can be effective against third parties as long as they are entered in a court order. *See* FED. R. EVID. 502(d); FED. R. EVID. 502(d) advisory committee's note. As is often the case in e-discovery, numerous aspects of the *Moore* and *Actos* protocols implicate serious concerns with the requesting party viewing or receiving privileged documents. However, aside from Part IV.E, an extensive discussion of the privilege concerns in predictive coding protocols is beyond the scope of this Comment. *See infra* Part IV.E.
127. *Moore*, 287 F.R.D. at 187.
128. *Id.* at 199.
129. *Id.*

relevant and irrelevant documents resulting from the search, subject to prompt return to MSL, and plaintiffs could give input on the relevance decisions made during the process.[130] The protocol required MSL to conduct seven total rounds of iterative training,[131] reviewing 500 documents between each round unless the change in the amount of relevant documents became less than 5% and no new "hot" documents were produced.[132] As part of the training process, the third-party software experts and MSL's lawyers were to work together in a "good faith effort" to select documents in the sample that would increase the coding software's accuracy.[133]

After the iterative training rounds but before the final search, the protocol required MSL to review 2,399 documents deemed irrelevant by the coding software to ensure the quality of the production.[134] This sample size provided a 95% confidence level with a ± 2% margin of error.[135] MSL turned these irrelevant

---

130. *Id.* at 202.

131. *Id.* at 187. However, while the court did endorse the seven rounds of iterative training by MSL, it also noted that more rounds might be necessary:

> But if you get to the seventh round and [plaintiffs] are saying that the computer is still doing weird things, it's not stabilized, etc., we need to do another round or two, either you will agree to that or you will both come in with the appropriate [quality control] information and everything else and [may be ordered to] do another round or two or five or 500 or whatever it takes to stabilize the system.

Id. (citing Transcript of Feb. 8, 2012 Conference at 76–77, *Moore*, 287 F.R.D. 182 (No. 88)).

132. *Id.* at 201–02. The term "hot" document is another way of saying "highly relevant" document. *Id.* The court also refers to these documents as "smoking gun" documents. *Id.* at 189. While this protection was built into the protocol, the court specifically noted that "[p]laintiffs reserve the right, at all times, to challenge the accuracy and reliability of the predictive coding process and the right to apply to the [c]ourt for a review of the process." *Id.* at 202.

133. *Id.*

134. *Id.* The purpose of this quality control step was "to allow calculation of the approximate degree of recall and precision of the search and review process used." *Id.*

135. The 2,399-document sample used to create the seed set was based on a 95% confidence level. *Id.* at 186. When the parties subsequently used a 2,399-document sample, it was based on the same confidence level. The confidence level represents the percentage probability that the sample document collection is a true estimate of the amount of relevant documents in the entire corpus of documents. NELSON, *supra* note 11, at 13. So, based on this quality control measure, there was a 95% chance, with ± 2% margin of error, that the quality control sample manually reviewed by MSL was reflective of the entire body of irrelevant documents. The confidence level is not a measure of accuracy, but, rather, it "refers to our belief in the measurement's reliability." Herbert L. Roitblat, *On Some Selected Search Secrets*, INFO. DISCOVERY BLOG (Jan. 9, 2012, 10:35 PM), http://orcatec .blogspot.com/2012_01_01_archive.html.

documents over to the plaintiffs for additional review.[136] Finally, MSL was to conduct a final search using the predictive coding software and manually review all documents that the software found to be relevant.[137] If MSL's manual review also found the documents to be relevant and non-privileged, then MSL produced the documents to the plaintiffs per the document request.[138] One additional provision of the protocol provided that MSL did not have to conduct the final search and review until any of the plaintiffs' objections with the process were resolved by either the parties or the court.[139] This completed the predictive coding protocol.[140]

B. In re Actos (Pioglitazone) Products Liability Litigation

1. Case Background

In re Actos (Pioglitazone) Products Liability Litigation[141] is a multidistrict litigation in which multiple plaintiffs sued Takeda Pharmaceutical claiming personal injury from using defendant's product Actos, a prescription blood sugar medication.[142] At the time that this Comment went to print, nearly 300 cases were consolidated in the multidistrict litigation.[143] The predictive coding protocol was implemented to assist Takeda in searching and reviewing ESI for production.[144] Takeda followed the comparison method[145] and employed a private third-party vendor, Epiq Systems, to assist with the search and review processes.[146]

---

136. *Moore*, 287 F.R.D. at 202.
137. *Id.*
138. *Id.*
139. *Id.* at 202–03.
140. The protocol was, in fact, more extensive than this. One important provision noted that the plaintiffs agreed to pay for their involvement in the predictive coding process. *Id.* Other provisions such as the "Format of Production" and "Timing" were also included. *Id.* at 203–04. However, an in-depth discussion of these provisions is not necessary for the purposes of this Comment.
141. *In re* Actos (Pioglitazone) Prods. Liab. Litig., No. 6:11-md-2299, 2012 WL 7861249 (W.D. La. July 27, 2012) (Case Management Order: Protocol Relating to the Production of Electronically Stored Information).
142. Complaint, *In re* Actos (Pioglitazone) Prods. Liab. Litig., No. 6:11-md-2299, 2013 WL 3171766 (W.D. La. June 13, 2013). The litigation was ongoing while this Comment was written.
143. *In re* Actos (Pioglitazone) Prods. Liab. Litig., No. 6:11-md-2299 (W.D. La. Dec. 7, 2012) (Conditional Transfer Order).
144. *In re Actos*, 2012 WL 7861249, at *4.
145. *See supra* note 110 and accompanying text.
146. *In re Actos*, 2012 WL 7861249, at *4. Epiq uses Equivio's predictive coding software called Relevance. Again, this Comment does not endorse any particular predictive coding vendor or software. For more information about

2. The Protocol

To begin the predictive coding process, Epiq collected e-mails from four major custodians associated with the defendant, Takeda. Epiq pooled these e-mail documents together, along with regulatory documents provided by Takeda, to form a "sample collection population," which is a random sample purposed to represent the entire document population being searched.[147] No seed set was created, though the parties could agree to create a seed set later if they felt it was necessary.[148]

Next, the protocol required that the plaintiffs and Takeda each nominate three experts to train the predictive coding tool.[149] These "experts" training the coding tool were actually lawyers who were familiar with the case.[150] Plaintiffs' experts signed a nondisclosure and confidentiality agreement in which they agreed not to disclose information that would be "subject to withholding or redaction under the Protective Order."[151] The experts received training documents detailing how to use the software, and then they received technical training on the coding software and process. The experts determined the relevance of documents in the control and training sets.[152] Takeda's lawyers led the computer training and had access to the sample collection in its entirety, but they were to "work collaboratively with [p]laintiffs' counsel during the [a]ssessment and [t]raining phases."[153]

The next step in the protocol was the assessment phase in which documents were reviewed to create the control set.[154] The predictive coding software generated a 500-document random sample from the

---

Equivio's Relevance software, see EQUIVIO, http://www.equivio.com (last visited Oct. 27, 2013).

147. *In re Actos*, 2012 WL 7861249, at *4. The sample is random so that is representative and reflects the larger population. J. DeLayne Stroud, *Basic Sampling Strategies: Sample vs. Population Data*, ISIXSIGMA (Feb. 26, 2010), http://www.isixsigma.com/tools-templates/sampling-data/basic-sampling-strategies-sample-vs-population-data/. Using a random sample helps reduce bias. *Id.*

148. *In re Actos*, 2012 WL 7861249, at *5.

149. *Id.* at *4.

150. *EDI Survey*, *supra* note 21, at 18.

151. *In re Actos*, 2012 WL 7861249, at *4. The court issued the protective order on July 30, 2012. *In re* Actos (Pioglitazone-Prods. Liab. Litig.), No. 6-11-md-2299, 2012 WL 3899669 (W.D. La. July 30, 2012) (Case Management Order: Protecting the Confidentiality of Discovery Materials).

152. *In re Actos*, 2012 WL 7861249, at *4. The control and training sets are discussed further *infra* Part III.B, IV.B.

153. *In re Actos*, 2012 WL 7861249, at *4.

154. *Id.* at *5.

sample collection population.[155] The parties' experts worked together to review and determine the relevance of the documents contained in the random sample.[156] These assessment-phase documents made up the control set, which functioned as a reference point for accuracy and helped determine "richness."[157] The assessment phase continued until the control set contained at least 385 relevant documents.[158]

When the parties completed the assessment phase, the iterative training phase began.[159] The coding software selected a random sample that consisted of 40 documents, and the experts from both parties worked together to determine the relevance of each document, then used that relevance information to train the software.[160] After each round, the software calculated its training status as either "Not Stable, Nearly Stable, or Stable."[161] The experts continued this training process (sampling and reviewing 40 documents between each training round) until the system was stable.[162] After each round of training, the coding tool used "Active Learning" to enhance the quality of the search.[163]

Before the final search for production, the parties conducted a quality control test by reviewing a 500-document sample that the

---

155. *Id.*

156. *Id.*

157. *Id.* at *5–6. The term "richness" means the "percentage of relevant documents in a population." *Id.* In this case, the parties agreed to a richness confidence level of 95%. *Id.* at *6.

158. *Id.* The 385-document requirement was not arbitrary. The purpose of this number was that it "yield[ed] an error margin on recall estimates of ± 5%." *Id.*

159. *Id. See supra* note 126 and accompanying text for the purpose of iterative rounds.

160. *In re Actos*, 2012 WL 7861249, at *6.

161. *Id.* at *7. The stability classification essentially describes the coding tool's ability to accurately classify documents. The system is deemed stable when the subsequent samples do not contribute anything, or at least very little, to the coding software's ability to classify documents. *Id.*

162. *Id.*

163. *Id.* "Active Learning" means "that each training sample is selected based on what has been learned from previous samples." *Id.* The Active Learning approach allows the coding tool to predict the relevance of a document based on previous determinations by the human coder. It "learns" from the previous human determinations, and based on what the system has learned, it chooses the next documents for the human to review in order to enhance its future learning. TOM GROOM, THREE METHODS FOR EDISCOVERY DOCUMENT PRIORITIZATION: COMPARING AND CONTRASTING KEYWORD SEARCH WITH CONCEPT BASED AND SUPPORT VECTOR BASED "TECHNOLOGY ASSISTED REVIEW-PREDICTIVE CODING" PLATFORMS 4 (2012), *available at* http://documents.jdsupra.com/af05a22e-b3f2-40f3-aa84-a3297099fa6f.pdf. The Active Learning approach maximizes the use of the information input from human-reviewed samples between iterative rounds. *In re Actos*, 2012 WL 7861249, at *7.

software classified as irrelevant—in other words, documents below the agreed-upon relevancy cutoff point—which gave the parties a 95% confidence level with a variable margin of error.[164] This assured the parties that the documents the coding tool classified as irrelevant did not contain highly relevant information and that the cutoff point was proportional.[165] When the software was trained and quality control was completed, the parties conducted the final search for production.[166] The protocol left it up to the parties to determine an adequate and proportionate relevancy cutoff point.[167] Takeda manually reviewed all documents above this cutoff point and produced the documents to the plaintiffs.[168] The parties also agreed to meet and review a sample of the documents that were classified above the agreed cutoff score but found irrelevant by the manual review.[169] The protocol established that Takeda did not have to conduct the final search until any objections brought by the parties were either resolved by the parties or by court adjudication.[170] This concluded the *Actos* predictive coding protocol.[171]

The protocols in *Moore* and *Actos* demonstrate just a few of the important choices that have to be made throughout the predictive coding process. Predictive coding is a complicated endeavor that can be implemented in a variety of ways. While the technical aspects of the process are vital to its success, it is also important that the parties approach it with a cooperative mindset.

III. PREDICTIVE PROBLEMS: DIFFERENCES IN THE PROTOCOLS

As *Moore* and *Actos* took very different paths toward implementing predictive coding protocols,[172] this Part addresses

---

164. *In re* Actos, 2012 WL 7861249, at *7. The protocol referred to this process as "Test the Rest." *Id.* For an explanation of confidence levels, see *supra* note 135 and accompanying text. The protocol describes the variable margin of error as follows: "The margin of error depends on the percentage of relevant documents in the Rest. For example, if 5% of the Rest documents are found to be relevant, the margin of error is 1.9%. If 1% are relevant, the margin of error is 0.8%." *In re Actos*, 2012 WL 7861249, at *7.

165. *In re* Actos, 2012 WL 7861249, at *7. For more discussion on proportionality in discovery, see *supra* note 80.

166. *In re* Actos, 2012 WL 7861249, at *8.

167. *Id.*

168. *Id.*

169. *Id.*

170. *Id.*

171. Similar to the protocol in *Moore*, the protocol also required that the plaintiffs pay for their involvement in the process and specified the format of production. *Id.* at *9.

172. *See supra* Part II.

whether both protocols adequately implemented the cooperative approach emphasized in e-discovery. First, it examines the use of experts in the protocols. Then, it looks at the advantages or disadvantages of seed sets and random samples. Next, it compares the parties' usage of iterative rounds and focuses on the quality control portions of the protocols. Finally, this Part concludes by assessing the final production of documents in each case.

A. A Lack of Expert Cooperation

The most glaring difference between the protocols in *Moore* and *Actos* was the use of experts. In *Moore*, defendant MSL's lawyers worked in good faith with a third-party vendor's expert to ensure that the software was adequately trained.[173] The plaintiffs did not play an active role in the process and only participated after the efforts of MSL and the third-party experts, though they were allowed to give input during the process and review the documents after the fact.[174] This process is antithetical to the cooperative and transparent principles of e-discovery.

First, most of the cooperation in the coding process occurred between MSL and the third-party vendor's experts in their "good faith" effort.[175] This negatively impacted cooperation because the *Moore* plaintiffs disputed certain aspects of MSL's implementation of predictive coding and brought their complaints to the court for resolution.[176]

Second, the plaintiffs did not have anyone representing their interests during the actual coding process, which made predictive coding a "black box" for the plaintiffs.[177] In other words, the process was not transparent for the plaintiffs and contributed to their need for "clarification."[178] The lack of transparency led to a decrease in

---

173. *See* discussion *supra* Part II.A.2.

174. *See* discussion *supra* Part II.A.2.

175. *See* discussion *supra* Part II.A.2.

176. *See* discussion *supra* Part II.A.1. It is always possible that confusion stemming from an e-discovery search method could arise from a lawyer's lack of experience with it. Whether this was the case for the plaintiff in *Moore* is irrelevant. Because predictive coding is so new to the courtroom, the defendant should have anticipated the need to show the search method's reasonableness. *See supra* note 81 and accompanying text (explaining "reasonableness").

177. "Black box" refers to the technical nature of the coding technology that can make it difficult to explain or understand. *Advice from Counsel*, *supra* note 17, at 6–7.

178. *See* discussion *supra* Part II.A.1. It is true that there is no discovery, ethical, or moral rule requiring a responding party to educate the requesting party on its search techniques. However, it may still be in the responding party's best interest to do so. For instance, in *Moore*, the plaintiffs wanted clarification on the

efficiency because the parties had to spend time and money in court instead of resolving the issue before discovery commenced or even during the discovery search.

The use of experts in *Actos* was drastically different than in *Moore*. The *Actos* parties each chose three experts to train the predictive coding system.[179] Each party paid for its involvement in the process, including the expert expenses.[180] The six experts reviewed training documents and received hands-on training regarding the use of the coding software.[181] Additionally, Takeda, the responding party, like MSL in *Moore*, led the process and had access to the entire corpus of documents.[182]

The *Actos* protocol promotes cooperation between the parties. The parties selected their own experts and were equally represented during the coding process. In *Moore*, the plaintiffs were forced to sit quietly and wait for the results of the search. But in *Actos*, both parties were represented and made decisions on document relevance during the process. The *Actos* approach eliminates the adversarial relationship between opposing parties and replaces it with a cooperative effort between experts.

At first glance, the *Actos* approach seems like a utopian collaborative and cooperative effort. The *Cooperation Proclamation* suggests that opposing parties work together to develop search strategies and methodologies,[183] but the *Actos* strategy actually surpasses this suggestion by requiring the parties to work together *during* the search process. Representation of both parties during the process also gives them a better chance to resolve disputes at the outset of discovery instead of going to the court after the fact.[184]

However, the expert strategy used in *Actos* may not be the utopian cooperative effort it seems to be and could give rise to problems. In the *Actos* approach, the parties both have three experts,

---

process. *See* discussion *supra* Part II.A.1. Clarification at the beginning of the process might have avoided the conflict altogether. *See* discussion *supra* Part I.B (explaining how transparency can help reduce discovery disputes). If a party wants "to secure the just, speedy, and inexpensive determination," then explaining the search technique to the opposing party before or during the process can help achieve that. FED. R. CIV. P. 1. Additionally, if both parties are satisfied with the reasonableness of the search, it is likely that the court will be also. *See supra* note 31 and accompanying text.

179. *See* discussion *supra* Part II.B.2.

180. *See* discussion *supra* note 172 and accompanying text.

181. *See* discussion *supra* Part II.B.2.

182. *See* discussion *supra* Part II.B.2.

183. *See supra* note 71 and accompanying text.

184. *See* discussion *supra* Part I.B (discussing how transparency can help reduce motion practice).

and because each party pays for its own experts, this opens the door to potential disputes between the experts. It is only natural that experts might feel a sense of loyalty to the party compensating them and might feel obligated to make relevance decisions in that party's best interest. Because each side has an equal number of experts, there is no way to work out problems if those experts disagree about document relevance. If the experts dispute the relevance of certain documents, the parties' lawyers might have to intervene. If the lawyers cannot resolve the issue, they would have to take the dispute to the court for resolution. In other words, the six-expert design could open the door to unnecessary disputes that would have to be worked out in court.[185]

Still, the expert approach in *Actos* does a better job of promoting transparency between the parties than the *Moore* approach. One particular concern in *Moore* was the "black box" nature of predictive coding.[186] But when a party chooses its own experts, it effectively lifts "the hood on predictive coding."[187] The plaintiffs have no adversarial relationship with their experts, therefore they can have open and candid discussions with the experts and be continuously updated on the search. If the parties have questions about how the predictive coding process works or is conducted, they can simply ask their experts who are actually using the predictive coding tool. This transparency also decreases the chance that the requesting party will challenge the reasonableness of predictive coding, because its own experts are playing an important role in the search process.[188]

The *Actos* expert approach also increases efficiency because the cooperation and transparency provided by the experts make it less likely that the parties will need to spend time and money litigating disputes. However, another aspect of the *Actos* approach decreases efficiency. The experts had to review software manuals and receive training to effectively use the coding software.[189] This takes more

185. *See* discussion *supra* Part I.A (discussing how the cooperative approach mainly strives to eliminate "unnecessary disputes").

186. *See* discussion *supra* Part III.A.

187. *Advice from Counsel*, *supra* note 17, at 7.

188. Predictive coding and similar technologies are new to the judicial system. Whether the court finds the search reasonable depends on how well the responding party explains the technology to the court. John E. Davis & Wayne C. Matus, *Does Your Search Pass Judicial Scrutiny?*, N. Y. L.J. (Oct. 27, 2008), *available at* http://www.pillsburylaw.com/siteFiles/Publications/DoesYourSearchPassJudicial Scrutiny.pdf. When the requesting party has experts who are involved and understand the process, it is only logical to conclude that that party also thinks the process is reasonable, unless the party specifically objects in light of consulting its experts.

189. *See* discussion *supra* Part II.B.2.

time and money than simply using an expert from the vendor who already knows how to use the software, as the defendants did in *Moore*. Additionally, the *Actos* approach requires parties to pay for their involvement in the protocol, and employing, training, and using the experts increases expenses.[190] The potential for these increased expenses is exacerbated by some of the complex tasks assigned to the experts, such as creating seed sets and control sets.

B. Seed Set Versus Control Set

In *Moore*, MSL started the coding process by generating a random sample of 2,399 documents using the predictive coding software.[191] Next, it created a seed set using keyword searches.[192] Keyword searches are considered reasonable in many cases.[193] In fact, they have become the "status quo" for e-discovery.[194] Nevertheless, they have also received heated, and perhaps well-deserved, judicial and scholarly criticism.[195] Keyword searches have been described as a "model of inefficiency"[196] and can cause disputes between parties, which hamper the cooperative effort.[197] Litigating disputes over keywords necessarily reduces efficiency by adding additional time and costs to e-discovery. Even in *Moore*, the plaintiffs had to supplement the keywords used by MSL to create the seed set, and MSL had to search again using the supplemental

---

190. *See* discussion *supra* note 172 and accompanying text.
191. *See* discussion *supra* Part II.A.2.
192. *See* discussion *supra* Part II.A.2.
193. *Sedona Conference Best Practices*, *supra* note 11, at 196.
194. Paul & Baron, *supra* note 1, at 37.
195. *See, e.g.*, U.S. v. O'Keefe, 537 F. Supp. 2d 14, 24 (D.C. 2008) (noting that "for lawyers and judges to dare opine that a certain search term or terms would be more likely to produce information than the terms that were used is truly to go where angels fear to tread"). Keyword searches have inherent practical problems. One of the most glaring problems is the "ambiguity and indeterminacy in human language." Paul & Baron, *supra* note 1, at 38. This language barrier can make finding responsive information very difficult. *Id.* One e-discovery author has equated keyword searches to a game of "Go Fish" because the requesting party, who does not know what "cards" the responding party holds, is essentially guessing which keywords may produce relevant information. *Go Fish*, *supra* note 90. Even more, the accuracy of keyword searches is suspect. It is estimated that these searches are only 22% to 57% effective when used to find responsive discovery documents. Robert C. Manlowe et al., *Paradigm Shifts in E-Discovery Litigation: Cooperate or Continue to Pay Dearly*, 78 DEF. COUNS. J. 170, 171 (2011).
196. *Go Fish*, *supra* note 90.
197. *See, e.g.*, *O'Keefe*, 537 F. Supp. 2d at 23–24 (noting the defendant's protest to the search terms used by the government in its electronic search for documents).

keywords, review 4,000 documents resulting from the search, and give them to the plaintiffs for review.[198] This truly is a "model of inefficiency."[199]

The parties in *Actos* formed an initial random sample of 500 documents by collecting documents from Takeda custodians to eventually form the control set.[200] The assessment phase continued until the 500-document sample contained at least 385 relevant documents.[201] The *Actos* protocol removes the keyword searching strategy, which can be fraught with problems, and replaces it with collaborative document review between the experts. This makes it a cooperative and efficient process that removes keyword searches and reduces the likelihood of controversy between the parties.

C. Tallying up the Totals of Iterative Rounds

The *Moore* protocol required defendant MSL to complete seven total iterative rounds to stabilize the coding software.[202] The court qualified this by noting potential uncertainty with the software and essentially ordered MSL to do as many rounds as it took to stabilize the software.[203] Between each round, MSL reviewed 500 documents to determine the system's progress.[204] This means MSL was required to review and code a minimum of 3,500 documents during the iterative rounds and possibly more if the software was still doing "weird things."[205] Compare this to MSL's goal of producing only 40,000 total documents:[206] It was required to review almost 9% of that amount during the iterative rounds alone.[207] In fact, MSL had to review 2,399 documents for the initial sample, 4,000 documents for plaintiffs' supplemental keywords, 3,500 documents during iterative rounds, and 2,399 documents for quality control.[208] In total, MSL had to review 12,298 total documents during the process, which is 30% of its 40,000-document goal. MSL undoubtedly chose predictive coding to increase efficiency, but the number of documents it had to review during the process and the increased litigation that the search process caused made the predictive

198. *See* discussion *supra* Part II.A.2.
199. *Go Fish*, *supra* note 90.
200. *See* discussion *supra* Part II.B.2.
201. *See* discussion *supra* Part II.B.2.
202. *See* discussion *supra* Part II.A.2.
203. *See supra* note 131 and accompanying text.
204. *See* discussion *supra* Part II.A.2.
205. *See supra* note 131 and accompanying text.
206. *See supra* note 116 and accompanying text.
207. *See* discussion *supra* Part II.A.2.
208. *See* discussion *supra* Part II.A.2.

coding's efficiency questionable at best, at least in that case. Additionally, the imposition of seven iterative rounds is arbitrary. There is no reason given for this number in either the *Moore* opinion or protocol. Setting this number gives the parties one more thing to dispute and can only negatively affect cooperation and efficiency.

In *Actos*, the protocol did not suggest a specific number of training rounds. Instead, the coding tool selected a 40-document sample between rounds that the parties reviewed and used to train and stabilize the system.[209] Comparing this to the 3,500 documents that the parties in *Moore* had to review, the *Actos* parties would have had to conduct more than 87 iterative rounds to review that amount of documents. Also, because there was no set number of rounds and the parties' experts were working together during the process to stabilize the coding software, the number of rounds would not be disputed. The *Actos* approach is thus a cooperative, transparent, and efficient effort.

D. The Reliability of Quality Control Samples and Their Impact on Efficiency

Both protocols required the parties to conduct a quality control test by manually reviewing a random sample of documents deemed irrelevant (or below the cutoff point) by the coding software.[210] The *Moore* protocol required MSL to review 2,399 non-responsive documents, while the *Actos* protocol required Takeda to review 500 non-responsive documents.[211] In each case, the parties chose to use a 95% confidence level when reviewing the quality control sample. Confidence levels are not a measurement of accuracy; confidence levels represent the probability that the random sample accurately reflects the entire document collection being searched.[212] The sample size, therefore, is partially based on the size of the larger document collection being searched and the chosen confidence level.[213] In other words, the different sized quality control samples used in *Moore* and *Actos* were not due to the confidence levels used because in each case the confidence level was 95%.

In addition to the confidence level and the document population size, another factor that impacts sample size is the margin of error. The term "margin of error" refers to the "maximum likely difference

209. *See* discussion *supra* Part II.B.2.
210. *See* discussion *supra* Parts II.A.2, B.2.
211. *See* discussion *supra* Parts II.A.2, B.2.
212. *See supra* note 135 and accompanying text.
213. *See, e.g.*, RAOSOFT, http://www.raosoft.com/samplesize.html (last visited Oct. 9, 2012) (number of documents in the population is a variable in the equation for determining the desired sample size).

between a true [document] population value and a sample estimate of that value."[214] The plaintiffs in *Moore* opted for a ± 2% margin of error, while a variable margin of error was used in *Actos*. As the margin of error becomes smaller, the sample size becomes larger in order to more accurately reflect the true document population.

Stepping back from how samples are actually developed, there are conflicting positions as to the reliability of quality control sampling in general.[215] Ralph Losey, a well-respected e-discovery practitioner and scholar, points to empirical studies that propose that any confidence level higher than 66% is unreliable because humans ultimately review the documents and can have differing opinions about the importance of a particular document.[216] Declining to go to the 66% extreme, Losey does suggest that a 95% confidence level with ± 5% margin of error is more realistic than, say, a more demanding ± 2% margin of error.[217] However, others have taken issue with Losey's analysis, noting that it is based on document reviews with multiple reviewers.[218] Having only one human reviewer can eliminate the variability of opinions, therefore making confidence levels and margins of error reliable.[219] These conflicting positions have a broader impact than just the quality control stage, though, as they could also affect how the final document production is coordinated.

E. Options for Reviewing Documents Prior to Final Document Production

Although creating a quality control sample involves some complex considerations, the final production of documents is a relatively simple decision, independent from the rest of the protocol.[220] Essentially, responding parties must decide how many documents, if any, they should manually review before producing those documents to the requesting party. The responding party has three options.[221] The first option is to produce the documents

---

214. NELSON, *supra* note 11, at 13.

215. *Compare* Ralph C. Losey, *Secrets of Search – Part III*, E-DISCOVERY TEAM (Dec. 29, 2011, 8:54 PM), http://e-discoveryteam.com/2011/12/29/secrets-of-search-part-iii/ [hereinafter *Secrets of Search*] ("[R]andom samples with 95% confidence levels ± 2 are also unrealistically high."), *with* Roitblat, *supra* note 135 ("[E]stimates can be made precise enough with reasonable sample sizes.").

216. *Secrets of Search*, *supra* note 215.

217. *Id.* Additionally, Losey describes aspirations of a 99% confidence level as "delusional." *Id.*

218. Roitblat, *supra* note 135.

219. *Id.*

220. NELSON, *supra* note 11, at 27.

221. *See generally id.* at 27–28.

deemed relevant by the predictive coding tool without manually reviewing any of them.[222] This is the most efficient option because no time or money is spent manually reviewing documents.[223] The second option is to manually review a random sample, which is also known as "spot-checking."[224] This approach essentially follows the same logic as the sampling strategy seen in the quality control check.[225] The third option requires the parties to manually review all of the relevant documents before producing them to the requesting party.[226] This approach is the least efficient of the three because additional time and money are expended to manually review each document.[227]

The responding parties in *Actos* and *Moore* both chose the third option, that is, to manually review all documents before producing them.[228] It is the least efficient document production option, but there are valid reasons for choosing it. The newness of predictive coding and parties' and courts' lack of experience with it is particularly concerning, and manually reviewing all relevant documents is one way to alleviate those concerns.[229] In other words, manually reviewing all documents makes the process more transparent, at least for the producing party. Still, a producing party might also choose to review all the documents so that it will know the content of the documents it is handing over to its adversary, which can help with case evaluation.[230]

Notably, the parties in *Actos* went a step further than just reviewing all documents. The protocol describes that the parties were to meet and review a sample of the documents that were above the agreed cutoff score but still found irrelevant by Takeda's manual review.[231] This step can be justified by the same transparency

---

222. *Id.* at 27.
223. *Id.*
224. *Id.* at 28.
225. *Compare id.* and accompanying text *with* discussion *supra* Part III.D.
226. NELSON, *supra* note 11, at 28.
227. *Id.*
228. *See* discussion *supra* Parts II.A.2, B.2.
229. *See* NELSON, *supra* note 11, at 28.
230. *See Evidence Mounting In The Case For Predictive Coding*, METROPOLITAN CORP. COUNS. (Sept. 22, 2012, 10:02 AM), http://www.metrocorpcounsel.com /articles/20599/evidence-mounting-case-predictive-coding (explaining that predictive coding can help "determine the strength of the case"). *But see* Joshua L. Fuchs & Benjamin J. Wolinsky, *Understand Predictive Coding Options*, TEX. LAWYER, Sept. 3, 2012, *available at* http://www.jonesday.com/files/Publication/b9673a23-4a8a-43e9-8272-bd3d5a621a0b/Presentation/PublicationAttachment/ef69d4cd-fbfb-4946-9b34-bf7ff4f6cab2/651091201%20Jones%20Day.pdf (explaining that not manually reviewing documents before production causes a loss in "important litigation team knowledge").
231. *See* discussion *supra* Part II.B.2.

reasoning as choosing to review all relevant documents. However, it also has the same drawback. Because parties must review more documents, all of which have been reviewed already by Takeda, this procedure increases discovery costs and decreases efficiency.

It is apparent that neither the *Actos* nor the *Moore* protocol is perfect—Protocols are made by humans, so no protocol will ever be perfect. However, by using parts of each protocol along with ideals from the cooperative approach, the predictive coding protocol can be improved and crafted to avoid expensive and unnecessary e-discovery battles.

IV. A COOPERATIVE PREDICTIVE CODING PROTOCOL

It should be noted that some of the differences between the *Moore* and *Actos* protocols are undoubtedly attributable to the capabilities or requirements of the software vendors that the parties chose to use. Still, some of the strengths and deficiencies in the protocols are a result of the choices that the parties made when creating them. Parties and courts should employ search strategies that encourage cooperation and transparency between the parties while also maximizing efficiency during the search process. This Part introduces a model protocol that maximizes the implementation of cooperation, transparency, and efficiency into each part of the protocol discussed in Part III.

A. Odd Man In: A Cooperative Expert Approach

The use of experts in both *Moore* and *Actos* was imperfect. In *Moore*, the cooperative effort between the parties was minimized, and the lack of transparency made the predictive coding process a "black box" technology for the plaintiffs.[232] In *Actos*, using three experts on each side maximized cooperation and transparency between the parties, but training the experts made the strategy less efficient because of increases in time and money.[233] Also, an even number of experts from both the plaintiffs and defendants raised the potential for disputes during the coding process.[234]

Nevertheless, party-specific experts can facilitate the predictive coding process. Because the experts make the search more transparent, they can help parties avoid disputes that might

232. *See* discussion *supra* Part III.A.
233. *See* discussion *supra* Part III.A.
234. *See* discussion *supra* Part III.A.

otherwise be taken to court for resolution.[235] However, the potential for unnecessary disagreements among experts that the court might have to resolve looms large. Therefore, this Comment suggests that there be an odd number of experts used during the predictive coding process.

For example, in *Actos* the parties could reduce the number of total experts from six to five. To do this, each party nominates two experts of their choosing for a total of four experts. The fifth expert would be chosen by both parties and paid equally by both parties.[236] This enhances the process in several ways. First, it increases efficiency by eliminating half of the cost associated with one expert because both parties are compensating the fifth expert. Second, because both parties are compensating the expert, he or she is able to remain unbiased when making decisions during the predictive coding process. Third, an odd number of experts allows for a "tie-breaker" if disagreements arise during the coding process, therefore resolving the problem of each expert siding with the party that employed him or her. Finally, and most importantly, this approach ensures the cooperative and transparent effort that was clearly desired in *Actos*.[237]

---

235. *See* discussion *supra* Part III.A. Additionally, experts save parties money by "facilitating the collection and review process," confining ESI into smaller parameters, and explaining search techniques to the court. Damian Vargas, *Electronic Discovery: 2006 Amendments to the Federal Rules of Civil Procedure*, 34 RUTGERS COMPUTER & TECH. L.J. 396, 416 (2008).

236. If the parties cannot agree on who will be the neutral expert, they are not without recourse. For example, one solution to this problem is for the parties to request the court to appoint a special master to be the neutral expert. *See* FED. R. CIV. P. 53. The parties might also get the court to appoint its own expert. *See* FED. R. EVID. 706. Additionally, parties should follow this model depending on the needs of the case. For instance, in smaller or less complex litigation, the parties might consider only three total experts, but in larger ESI cases, they might employ five or seven total experts. This flexible case-by-case approach allows parties to appropriately balance costs while maintaining the cooperative effort.

237. While this use of experts may be appropriate in specific cases, differing protocols are sometimes due to differing situations. *See* NELSON, *supra* note 11, at 14 (providing reasons why predictive coding workflows might differ). For instance, *Actos* was a multidistrict litigation, while *Moore* was simply a multiple-plaintiff litigation. This distinguishing factor can dramatically impact the plaintiffs' approaches in discovery. Because there are usually a large number of plaintiffs in multidistrict litigation, they can combine their resources to make the case less expensive to litigate. Kathleen Michon, *Multidistrict Litigation (MDL) for Drug Lawsuits and Other Cases*, NOLO, http://www.nolo.com/legal-encyclopedia/multidistrict-litigation-mdl-drug-lawsuits-32952.html (last visited Oct. 27, 2013). In other words, paying experts to facilitate the predictive coding process would not financially impact the plaintiffs in *Actos* as much as it might in a case with only five plaintiffs like *Moore*. Nevertheless, the plaintiffs in *Moore* did have to spend additional money objecting to the predictive coding protocol.

B. The Control Set and "Keyword-Free" Coding

Moore brought attention to some of the problems with seed sets.[238] Because seed sets involve keyword searches, they can create problems that seriously strain the cooperative effort between opposing parties.[239] The use of seed sets may have another critical downside: It may actually bias the predictive coding process.[240] Instead of training the software to find documents based on their content, lawyers using seed sets are able to train the software to find documents that give them "self-reinforcing results."[241] In other words, seed sets can skew the training process because they only train the coding tool to incorporate what the coder already knows.[242]

Actos started the process from scratch and developed a control set instead of seed sets.[243] Predictive coding protocols should follow the *Actos* model for two reasons. First, eliminating keyword searches also eliminates the possibility of keyword disputes, which increases cooperation between the parties. Second, it eliminates the possibility of attorney bias and creates a higher quality search because the coding software is able to develop its own idea of document relevance.[244] Overall, the *Actos* strategy's incorporation of a keyword-free control set, instead of seed sets, encourages cooperation, increases efficiency, and enhances the quality of the search. Future protocols should follow its lead.

See discussion *supra* Part III.A. It is possible that the increased financial burden resulting from employing experts, who can increase cooperation and transparency and potentially avoid disputes, might have reduced or eliminated the costs of litigating the discovery dispute.

238. *See* discussion *supra* Part II.A.2.

239. *See supra* note 195 and accompanying text.

240. John Tredennick, *Judge Peck Provides a Primer on Computer-Assisted Review*, CATALYST E-DISCOVERY SEARCH BLOG (Mar. 14, 2012), http://www.catalystsecure.com/blog/2012/03/judge-peck-provides-a-primer-on-computer-assisted-review/.

241. *Id.* This is essentially the same issue addressed in *O'Keefe*, that is, that the electronic search is only as good as the responding lawyer's thoughts. *See supra* note 198 and accompanying text.

242. Greg Buckles, *Actos Case TAR Protocol Order – Equivio's Relevance in Action?*, EDISCOVERYJOURNAL (Aug. 14, 2012, 3:58 PM), http://ediscoveryjournal.com/2012/08/actos-case-tar-protocol-order-equivios-relevance-in-action.

243. *See* discussion *supra* Part II.B.2.

244. Tredennick, *supra* note 240. *See also* discussion *supra* Part II.B.2.

C. Increasing Cooperation and Efficiency by Eliminating Arbitrary Iterative Rounds

Future protocols should again look to *Actos* when conducting iterative rounds. In *Actos*, the parties used Active Learning and reviewed a 40-document sample between an indefinite number of rounds, while the *Moore* protocol required review of a 500-document sample between seven iterative rounds.[245] The *Actos* strategy is more efficient than the *Moore* strategy because the experts reviewed fewer documents between training rounds.[246] Additionally, protocols should never apply an arbitrary number of iterative rounds. The purpose of the iterative training process is to stabilize the coding software.[247] Until this is accomplished, the software's search will be neither effective nor reliable. Setting an arbitrary number of iterative rounds to train the software only gives the parties another opportunity to disagree and could hinder the cooperative effort.[248]

D. Implementing a Larger Margin of Error into Quality Control Sampling

As previously discussed, the total size of the document collection, along with the confidence level and margin of error employed by the parties, determines the amount of documents subject to manual review during the quality control stage.[249] Ultimately, parties have to decide if they should put their faith in a confidence level and margin of error. This choice hinges on which

---

245. *See* discussion *supra* Part III.C.

246. *See* discussion *supra* Part III.C. The difference in the number of documents reviewed between rounds in each case was likely attributable to the Active Learning technology that was used in *Actos*. *See supra* note 163 and accompanying text. Essentially, using the Active Learning technology made the training process much more efficient and arguably more accurate. The plaintiffs in *Moore* coded the seed set, which was created by keywords. *See* discussion *supra* Part II.A.2. The expert coders knew the responsiveness of the documents as they were coded. On the other hand, in the Active Learning approach, the coding tool, as part of iterative training, selects documents that it is most uncertain as to their responsiveness. Jeremy Pickens, *The Recommind Patent and the Need to Better Define 'Predictive Coding'*, CATALYST E-DISCOVERY SEARCH BLOG (June 13, 2011), http://www.catalystsecure.com/blog/2011/06/the-recommind-patent-and-the-need-to-better-define-predictive-coding/. In other words, Active Learning seeks to learn unknown information instead of being coded known information.

247. *See supra* note 126 and accompanying text.

248. *See* discussion *supra* Part III.C.

249. *See* discussion *supra* Part III.D.

position the parties take toward the reliability of confidence levels.[250]

As mentioned earlier, there are two positions concerning the reliability of confidence levels and margins of error. Losey suggests a modest change from a ± 2% to ± 5% margin of error based on information that implies that these calculations may not be completely reliable.[251] This makes the process more efficient—A larger margin of error dictates a smaller sample size to manually review. The countering position suggests having only one person manually review the documents. This approach eliminates the variable of multiple human opinions, therefore supposedly making confidence levels more reliable.[252]

Practically speaking, having only one person reviewing documents would likely be undesirable for most litigants. Both *Moore* and *Actos* were "high-stakes" litigation. This type of litigation often involves multiple lawyers on each side, and because not all lawyers think alike, one lawyer might be able to form a litigation strategy from a document that another lawyer would consider irrelevant. Therefore, although having only one person reviewing documents would allow for a definitive relevance determination, it could limit the quality of representation in high-stakes or important civil rights cases.

Alternatively, increasing the margin of error, as Losey suggests, has a tremendous impact on the efficiency of the manual review of the samples. For instance, in *Moore*, the parties agreed to a 95% confidence level with a ± 2% margin of error, which required manual review of 2,399 documents.[253] Keeping the 95% confidence level and changing the margin of error to ± 5% reduces the amount of documents subject to manual review to 385 documents.[254] This eliminates more than 2,000 documents subject to manual review. Because there is no evidence that demonstrates substantial indicia of reliability of document sampling and confidence levels, parties should modestly increase the margin of error from ± 2% to ± 5% to increase the efficiency of quality control.

---

250. *See* discussion *supra* Part III.D.
251. *See* discussion *supra* Part III.D.
252. *See* discussion *supra* Part III.D.
253. *See* discussion *supra* Part II.A.2.
254. *See* RAOSOFT, *supra* note 213. This number was calculated by using the statistics from *Moore*: a 5% margin of error, a 95% confidence level, a 3,000,000-document population size, and a 50% response distribution.

E. Spot-Checking Before Final Document Production

The *Moore* and *Actos* protocols both required the responding party to manually review all documents before producing them. The need for transparency due to the newness of predictive coding might have justified that decision in both cases.[255] Manually reviewing all documents before production also allows the producing party to make a more accurate case assessment.[256] Nevertheless, manually reviewing all documents prior to production decreases efficiency and can add significant burdens and expenses to the process.[257]

Deciding which production review strategy to use should be made on a "case-by-case basis,"[258] but the best strategy is to spot-check the documents before production. Technology is not always reliable, so ensuring the quality of production is desirable to a certain extent. Therefore, manually reviewing at least some documents is necessary. Spot-checking a document collection maintains a certain level of transparency and allows the producing party to glimpse what it is producing, while also minimizing the amount of documents manually reviewed and making the process as efficient and cost-effective as possible. Additionally, the parties should use a higher confidence level and smaller margin of error for this sample than they would for the quality control sample.[259] This increases the amount of documents that the responding party reviews before production, which is desirable because the producing party will have a better idea of what it is turning over to its adversary.[260] Reviewing a larger sample with a higher confidence level or lower margin of error at this stage is still much more efficient than reviewing all of the relevant documents before producing them to the opposing party.

---

255. *See* discussion *supra* Part III.E.

256. *See supra* note 230 and accompanying text.

257. *Contra Achieving Quality*, *supra* note 75, at 321 (discussing how spot-checking documents before production can decrease the burdens and costs associated with manually reviewing all documents).

258. NELSON, *supra* note 11, at 27.

259. For example, the parties in *Moore* and *Actos* used a 95% confidence level for the quality control samples. In *Moore*, the margin of error was ± 2%, while the *Actos* margin of error was variable. *See* discussion *supra* Part II.A.2, B.2.

260. The confidence level and margin of error are determining factors in calculating the size of the document sample. *See* discussion *supra* Part IV.D. For instance, the parties in *Moore* reviewed a 2,399-document sample using a 95% confidence level with a ± 2% margin of error, but increasing the margin of error to ± 5% results in a 385-document sample. *See* discussion *supra* Part IV.D. Thus, using a higher confidence level or smaller margin of error would increase the sample size. However, parties would still review considerably fewer documents in a larger sample than if they reviewed all of the documents prior to production.

One important caveat to this production method is that it may raise significant privilege and confidentiality concerns. Because the producing party is sampling the documents, it will not check every document for privileged and confidential information. This makes it likely that privileged information will be produced to the requesting party.[261] Therefore, it is imperative that parties have proper privilege protection, such as clawback agreements, established in a court order.[262] Producing lawyers should also consult clients to assess their interest concerning the cost savings of spot-checking the production documents versus the potential for turning over privileged or confidential information.

Finally, the *Actos* protocol required an additional step: The parties were to meet and manually review a document sample that was deemed relevant by the coding tool but deemed irrelevant by Takeda's manual review.[263] This step is overkill. These specific documents went through three phases of review. They were reviewed by the predictive coding tool, manually reviewed by Takeda, and reviewed again by both parties after they were found to be irrelevant. This step may provide increased transparency and peace of mind, but reviewing the same documents three times is extremely inefficient and unnecessary.[264]

---

261. With large amounts of ESI, this inadvertent disclosure is virtually inevitable. Dennis R. Kiker, *Waiving the Privilege in a Storm of Data: An Argument for Uniformity and Rationality in Dealing with the Inadvertent Disclosure of Privileged Materials in the Age of Electronically Stored Information*, 12 RICH. J. L. & TECH. 15, *2 (2006). However, there are measures that the producing party can take to minimize the chances of inadvertent disclosure of privileged information. One of the most sensible is to use the coding tool before final production by coding "privilege vocabulary." *Achieving Quality*, *supra* note 75, at 319. This step, in addition to sampling, can minimize the amount of privileged information that will be disclosed.

262. *See supra* note 126 and accompanying text.

263. *See* discussion *supra* Part II.B.2.

264. It is also questionable if these documents, which have been determined irrelevant by a prior review, are within the scope of discovery. While making sure the irrelevant documents are, in fact, irrelevant could be "reasonably calculated to lead to the discovery of admissible evidence," it could also be argued that the documents are not "relevant" to the requesting party's claim because the documents have been deemed to be non-responsive. FED. R. CIV. P. 26(b)(1). Additionally, while transparency in the discovery process is encouraged, it can be a "slippery slope":

> The responding party has a right to privacy. They should not be required to give the requesting party the keys to the server room, the whole deck of cards. The requesting party is either suing the responding party, or being sued by the responding party. Either way, the requesting party should not be permitted to enter and search every nook and cranny of their adversary's inner sanctum.

Due to the newness of predictive coding in the legal system, it will likely take some time for parties to perfect the process. Nevertheless, using an odd number of experts, keyword-free control sets, iterative rounds that are not arbitrarily limited, a modestly increased quality control sample margin of error, and representative spot-checking or sampling prior to the final document production will help to ensure that predictive coding is an effective discovery tool. Equally important, these approaches will help to avoid unnecessary discovery battles and wasteful litigation by transforming a complex electronic search process into a cooperative, transparent, and efficient endeavor.

CONCLUSION

As the amount of ESI continues to increase, the legal system will need to evolve and find new ways to conduct effective and reasonable e-discovery. For now, predictive coding provides hope for parties seeking to reduce the costs and other problems associated with e-discovery. Because predictive coding is so new in the courtroom, it is not surprising that the implementation of the technology differed so greatly in *Moore* and *Actos*. However, it is essential that parties and courts make sure that the technology is used in the most cooperative, transparent, and efficient manner possible. When parties decide to use predictive coding, utilizing the model protocol proposed in this Comment will facilitate "just, speedy, and inexpensive" e-discovery.[265]

L. Casey Auttonberry[*]

Go Fish, *supra* note 90. This is particularly true because the coding software has deemed the documents as non-responsive to the needs of the case.

265. FED. R. CIV. P. 1.

* J.D./D.C.L., 2014, Paul M. Hebert Law Center, Louisiana State University. Special thanks to Professor Margaret S. Thomas and Professor Jeffrey C. Brooks for their guidance and encouragement while preparing this Comment.