Predictive Coding: Where Are We Now?

By: Ariana F. Pellegrino, Scott A. Petz, and Salina M. Hamilton[1]

Executive Summary
With hundreds of billions of electronic documents being exchanged every day, predictive coding and other technology-assisted review protocols continue to gain growing acceptance from practitioners and courts facing overwhelming amounts of data. Predictive coding combines human review with teachable algorithms to review and analyze large amounts of electronically-stored information without requiring attorneys or other reviewers to put eyes on every document. When used correctly, predictive coding is a game-changer for the discovery practice, saving substantial time, energy, and valuable resources.

Scott A. Petz is a Member in Dickinson Wright's Detroit office. Mr. Petz focuses his practice in the areas of commercial and business litigation, class and collective actions, labor litigation, consumer protection, and condemnation and land use. His email address is spetz@dickinson wright.com.

Salina M. Hamilton is an Associate in Dickinson Wright's Detroit office. Ms. Hamilton focuses her practice in the areas of commercial and business litigation. Her email address is shamilton@dickinson-wright.com.

Ariana F. Pellegrino is an Associate in Dickinson Wright's Detroit office. Mrs. Pellegrino focuses her practice in the area of commercial litigation. Her email address is apellegrino@dickinson-wright.com.

Introduction

In 2015 alone, the number of e-mails exchanged **every day** totaled over 205 billion and that figure is expected to grow to 246 billion e-mails per day by 2019.[2] Authorities estimate that upwards of 90 percent of all business documents are created electronically, with many of those documents never being reduced to paper form.[3]

The need to use discovery tools to efficiently and effectively search this ever-growing mass of electronically-generated data is evident, and predictive coding is rapidly gaining momentum as the tool-of-choice for large-scale discovery matters. Over four years ago, the Southern District of New York in *Moore v Publicis Groupe*[4] approved the use of predictive coding to search for relevant electronically-stored information ("ESI") in appropriate cases. Since coming to the forefront in *Moore*, the need to employ technologies like predictive coding—and judicial acceptance of such technologies—continues to grow.[5]

Predictive coding is here to stay, and those in the legal market should know what it is and when and how to use it.

Predictive Coding: What is it?

Predictive coding[6] is a technological tool that uses advanced algorithms in conjunction with human review to determine the relevancy of ESI.[7] In practice, using predictive coding during the document review process involves a series of steps that begins with subject-matter experts—typically the senior attorney and his or her core team—manually reviewing a sample set of documents, known as a "seed set."[8] The seed set is then used to train the predictive-coding software to recognize the properties of relevant documents, which are used to electronically code other documents.[9]

In short, the predictive-coding software will learn to evaluate the content, order, and arrangement of documents to identify and distinguish relevant documents from irrelevant documents, and it can do so in a way that goes far beyond simple keyword-searching. The review team then manually reviews these additional documents for accuracy, adding documents to the seed set to further enhance the software's capability and accuracy. In other words, predictive coding software is capable of iterative learning—meaning that the software continues to refine the coding of documents based upon input over time.

Eventually, after the manual reviewers' coding and the software's coding "sufficiently coincide," the software may be deemed to have learned enough to confidently predict the coding for any remaining documents.[10] To that end, through

the use of such technology, the review team "needs to only review a few thousand documents to train the computer,"[11] thereby potentially saving thousands of hours and resources that would otherwise need to be devoted to a full, manual review.

> [P]redictive coding is rapidly gaining momentum as the tool-of-choice for large-scale discovery matters.

Predictive Coding: When and How to Use It?

In the 2012 landmark decision in *Moore*, the Southern District of New York held that predictive coding is "an acceptable way to search for relevant ESI in appropriate cases."[12] Since *Moore*, predictive coding has largely been approved by those courts that have considered its use.[13]

The increased acceptance of predictive coding is largely unsurprising given that empirical data shows that predictive coding produces significantly better results than manual review—which is subject to human-error and inconsistency amongst review teams—at a fraction of the time and expense.[14] Notwithstanding, questions remain regarding the practical use of predictive coding.

Since *Moore*, courts have grappled with whether the use of keyword searches to narrow the pool of potentially relevant documents prior to the application of predictive-coding software is appropriate. Moreover, courts have considered to what degree litigants must engage in transparency and cooperation with respect to the use of the technology.

Use of Keyword Searches Prior to Using Predictive Coding

Courts generally agree that keyword searches may be performed before using the predictive-coding process. As one court explained, the practice of permitting parties to begin compiling responsive documents comports with the principle that "[r]esponding parties are best situated to evaluate the procedures, methodologies, and techniques appropriate for . . . producing their own electronically stored information."[15] However, keyword searches, standing alone, are likely insufficient to identify all responsive material. Instead, keyword searches should be combined with statistical sampling or other testing to ensure that responsive ESI has been sufficiently identified.

For example, in *Biomet*, the Northern District of Indiana held that a party's initial keyword search to narrow the pool of documents was appropriate and that the party would not be required to start over by using predictive coding on the whole pool.[16] In that case, before utilizing predictive coding to identify relevant ESI, Biomet narrowed the pool of documents from 19.5 million documents to 2.5 million documents by using keyword searches and removing duplicates.[17] Statistical sampling projected that only "between .55 and 1.33 percent" of the documents excluded as a result of keyword searching would be responsive.[18] Notwithstanding, the plaintiffs refused Biomet's offer to suggest additional keyword search terms and argued that the initial keyword search "tainted the process."[19]

> Courts generally agree that keyword searches may be performed before using the predictive-coding process.

Accordingly, the plaintiffs argued that Biomet should be required to repeat the initial search for responsive documents by applying predictive coding to all of the 19.5 million documents.[20] The court rejected the plaintiffs' position, holding that Biomet would not be required to "go back to Square One" where the large financial burden associated with uncovering a small number of responsive documents would violate the proportionality standard under the Federal Rules of Civil Procedure.[21]

> The increased acceptance of predictive coding is largely unsurprising given that empirical data shows that predictive coding produces significantly better results than manual review—which is subject to human-error and inconsistency amongst review teams—at a fraction of the time and expense.[14]

In contrast to *Biomet*, other courts have found keyword searching insufficient when applied alone because "the use of keywords without testing and refinement (or more sophisticated techniques) will in fact not be reasonably calculated to uncover all responsive material."[22] Given these contrasting principles, whether or not initial keyword searches will withstand scrutiny may depend on case-specific factors, such as the scope of the issues in the case and the scope of the potentially responsive ESI.

Cooperation and Transparency in Using Predictive Coding

While the use of predictive coding has been regularly approved by courts, case law in the post-*Moore* legal landscape generally stresses the importance of cooperation and transparency in its use.

For example, in *Bridgestone Americas, Inc. v. International Business Machines*

Corp,[23] the Middle District of Tennessee granted a party's request to use predictive-coding technology on documents that were previously identified using mutually agreed-upon search terms. The producing party, IBM, was "between one-third and one-half completed" with its manual review of the ESI when it requested approval to employ predictive coding.[24] The court recognized that the decision to permit predictive coding was a "judgment call" and permitted the use of the technology, in part because the producing party offered to provide its training documents to opposing counsel.[25] In its holding, the court reiterated the importance of communicating "on a frequent and open basis" regarding the production.[26]

> In contrast to *Biomet*, other courts have found keyword searching insufficient when applied alone because "the use of keywords without testing and refinement (or more sophisticated techniques) will in fact not be reasonably calculated to uncover all responsive material."[32]

Similarly, in *In re Actos*,[27] the Western District of Louisiana permitted the use of predictive coding where the predictive-coding protocol was sufficiently transparent. Specifically, the court ordered that the parties' experts would "have access to the entire sample collection population to be searched," that they would lead the computer training, and that they would identify privileged documents.[28] The court further ordered the parties to meet to review a random sample of documents for quality-control purposes.[29]

The role of cooperation and transparency with respect to seed sets is often a hot topic with parties. Indeed, *Moore* and its progeny left unanswered whether parties must produce their seed sets in order to satisfy the transparency requirement. While some courts avoid the need to answer this question where the producing party voluntarily discloses its seed set,[30] other courts have specifically rejected attempts to obtain the producing party's seed set, holding that the discoverability of such information is subject to the traditional limitations of relevancy under the Federal Rules of Civil Procedure.[31]

In *Biomet*, for example, the court held that a party's request for production of the seed set "reache[d] well beyond the scope of any permissible discovery by seeking irrelevant or privileged documents used to tell the algorithm what not to find."[32] Thus, where Biomet had produced all relevant and discoverable documents used in its seed set, the court held that the opposing party was not entitled to know "how [Biomet] went about identifying and selecting the documents . . . that it has produced" or "how Biomet used certain documents before disclosing them."[33]

Similarly, in *Freedman v Weatherford International Limited*, the Southern District of New York rejected a party's motion to compel the production of seed searches, holding that "discovery on discovery" would not remedy the perceived discovery defects.[34] In *Freedman*, the plaintiffs alleged improper tax practices in a class action against Weatherford.[35] Weatherford engaged an auditor to conduct an investigation of its earning statements and announced it would correct any errors in such statements. The plaintiffs, in turn, sought to compel the production of eighteen e-mails uncovered in the investigation along with the seed documents, even though they conceded that only three of the e-mails were likely to be responsive.[36] Plaintiffs additionally sought the production of a "report of the documents 'hit' by search terms used in connection with the . . . [investigation]."[37] The court denied the plaintiffs' request, holding that, while "[i]t is unsurprising that some relevant documents may have fallen through the cracks, … the Federal Rules of Civil Procedure do not require perfection."[38] Accordingly, because a significant percentage of the relevant documents would have already been identified by the contemplated searches, the plaintiffs were not entitled to additional discovery vis-à-vis the seed set.[39]

While courts have issued predictive-coding protocols requiring the producing party to disclose its seed set to opposing counsel,[40] Magistrate Judge Peck—who rendered the seminal decision in *Moore*—has explained that the production of a seed set is not the only means of ensuring transparency.[41] Rather, a party can show that predictive coding was used appropriately by several methods, including "statistical estimation of recall at the conclusion of the review," seeking "gaps in the production" and "quality control review of samples from the documents categorized as non-responsive."[42]

Conclusion

While the law on predictive coding is still developing, one thing is certain: predictive coding is here to stay. As Magistrate Judge Peck has proclaimed: "[I]t is black letter law that where the producing party wants to utilize [predictive coding] for document review, courts will permit it."[43]

Endnotes

1 The authors would like to thank Alma Sobo for her contributions to this article as a 2015 Dickinson Wright PLLC Summer Associate.

2 The Radicati Group, *E mail Statistics Report, 2015-2019*.

3 See Ronald J. Hedges, Daniel Riesel, Donald W. Stever & Kenneth J. Withers, *Taking Shape: E-Discovery Practices Under the Federal Rules*, SN085 ALI-ABA 289, 292 (2008); Robert M. Vercruysse, Gregory V. Murray, *Electronically Stored Information and the New Federal Rules of Civil Procedure Regarding Discovery* (2007), *available at* http://www.vmclaw.com/articles/3_Electronic_discovery.pdf; *NALA Manual for Paralegals and Legal Assistants* (6th ed. 2014), p. 321.

4 *Moore v Publicis Groupe*, 287 FRD 182 (SDNY, 2012) (Mag. J. Peck). Magistrate Judge Peck's opinion was adopted by Judge Andrew L. Carter, Jr. *Moore v Publicis Groupe*, unpublished opinion and order of the United States District Court for the Southern District of New York, issued April 25, 2012 (Docket No. 11-CV-1279), 2012 WL 1446534.

5 See, e.g., *Rio Tinto PLC v Vale SA*, 306 FRD 1245, 127 (SDNY, 2015); *Green v Am Modern Home Ins Co*, unpublished order of the United States District Court for the Western District of Arkansas, issued November 24, 2014 (Docket No. 1:14-CV-04074); 2014 WL 6668422; *Gabriel Technologies Corp v Qualcomm Inc*, unpublished order of the United States District Court for the Southern District of California, issued February 1, 2013 (Docket No. 08CV1992 AJB MDD); 2013 WL 410103; *Hinterberger v Catholic Health Sys, Inc*, unpublished decision and order of the United States District Court for the Western District of New York, issued May 21, 2013 (Docket No. 08-CV-380S F); 2013 WL 2250591; *Global Aerospace, Inc v Landow Aviation, LP*, unpublished order of the Circuit Court of Virginia, Louden County, issued April 23, 2012 (Docket No. CL 610040); 2012 WL 1431215; *In re Biomet M2a Magnum Hip Implant Products Liab Litig*, unpublished order of the United States District Court for the Northern District of Indiana, issued April 18, 2013 (Docket No. 3:12-MD-2391); 2013 WL 1729682, *2; *Kleen Products LLC v Packaging Corp of Am*, unpublished memorandum opinion and order of the United States District Court for the Northern District of Illinois, issued September 28, 2012 (Docket No. 10 C 5711); 2012 WL 4498465, *5; *In re Actos (Pioglitazone) Products Liab Litig*, unpublished order of the United States District Court for the Western District of Louisiana, issued July 27, 2012 (Docket No. 6:11-MD-2299); 2012 WL 7861249; *Bridgestone Americas, Inc v Int'l Bus Machines Corp*, unpublished order of the United States District Court for the Middle District of Tennessee, issued July 22, 2014 (Docket No. 3:13-1196); 2014 WL 4923014, *2; *but see EORHB, Inc v HOA Holdings LLC*, unpublished order of the Court of Chancery of Delaware, issued May 6, 2013 (Docket No. CIV.A 7409-VCL); 2013 WL 1960621, *1 (refusing to require the use of predictive coding where it "would likely be outweighed by any practical benefit" due to the "low volume of relevant documents expected to be produced").

6 Predictive coding technology is also commonly referred to as computer-assisted review ("CAR") or technology-assisted review ("TAR"). But see Paul Burns & Mindy Morton, *Technology-Assisted Review: The Judicial Pioneers*, Sedona Conf J, vol. 15, Fall 2014, *available at* http://www.americanbar.org/content/dam/aba/administrative/litigation/materials/2014_sac/2014_sac/technology_assisted_review_the_judicial_pioneers_authcheckdam.pdf (discussing the technical meaning and use of terms like predictive coding, CAR, and TAR).

7 See Scott A. Petz & Thomas D. Isaacs, *Predictive Coding: The ESI Tool Of The Future?*, Michigan Defense Quarterly, vol. 29, no. 1, July 2012, for an overview of predictive coding technology.

8 *Id.*

9 *Id.*

10 *Id.*, citing and quoting Andrew Peck, *Search, Forward*, L. Tech. News (Oct. 2011), http://www.law.com/jsp/lawtechnologynews//PubArticleLTN.jsp?id=1202516530534 (registration required to access).

11 *Id.*, quoting Andrew Peck, *Search, Forward*, L. Tech. News (Oct. 2011), http://www.law.com/jsp/lawtechnologynews//PubArticleLTN.jsp?id=1202516530534 (registration required to access).

12 *Moore*, 287 FRD at 183.

13 *See*, n 5 *supra*.

14 Maura R. Grossman & Gordon V. Cormack, *Technology-Assisted review in E-Discovery Can Be More Effective and More Efficient than Exhaustive Manual Review*, XVII Rich JL & Tech 11 (2011), http://jolt.richmond.edu/v17i3/article11.pdf

15 *Kleen Products*, 2012 WL 4498465 at *5, citing 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, 193 (Fall 2007).

16 *Biomet*, 2013 WL 1729682 at *2.

17 *Id.* at *1.

18 *Id.*

19 *Id.* at *2.

20 *Id.*

21 *Id.*

22 *Nat'l Day Laborer Org Network v US Immigration & Customs Enforcement Agency*, 877 F Supp 2d 87, 110 (SDNY, 2012).

23 *Bridgestone Americas, Inc v Int'l Bus Machines Corp.*, unpublished order of the United States District Court for the Middle District of Tennessee, issued July 22, 2014 (Docket No. 3:13-1196); 2014 WL 4923014.

24 *Id.* at *2.

25 *Id.*

26 *Id.*

27 *In re Actos*, 2012 WL 7861249.

28 *Id.* at *4.

29 *Id.* at *7.

30 See *Bridgestone.*, 2014 WL 4923014 at *2; see also *Rio Tinto PLC*, 306 FRD at 129.

31 See *Freedman v Weatherford Intern Ltd*, unpublished memorandum and order of the United States District Court for the Southern District of New York, issued September 12, 2014 (Docket No. 12 CIV. 2121-LAK-JCF); 2014 WL 4547039, *1; see also *Biomet*, 2013 WL 6405156 at *1.

32 *Biomet*, 2013 WL 6405156 at *1.

33 *Id.* at *1-2.

34 *Freedman*, 2014 WL 4547039 at *1.

35 *Id.*

36 *Id.* at *2.

37 *Id.* at *3.

38 *Id.*, quoting *Moore*, 287 FRD at 191.

39 *Id.*

40 See *Actos*, 2012 WL 7861249 at *4 (issuing a predictive coding protocol stating: "[a] ttorneys representing Takeda (a related pharmaceutical drug) will have access to the entire sample collection population to be searched and will lead the computer training, but they will work collaboratively with Plaintiffs' counsel during the Assessment and Training phases.").

41 See *Rio Tinto PLC*, 306 FRD at 128 (delineating procedures for evaluating documents in the event of non-cooperation by the producing party, including statistical estimation of recall and random sampling); see also *In re Lithium Ion Batteries Antitrust Litig*, unpublished order of the United States District Court for the Northern District of California, issued February 24, 2015 (Docket No. 13-MD-02420-YGR (DMR)); 2015 WL 833681, *2 (holding that random sampling of documents is the best way to refine searches and improve precision).

42 *Id.*, citing Maura R. Grossman & Gordin V. Cormack, *Comments on "The Implications of Rule 26(g) on the Use of Technology-Assisted Review,"* 7 Fed Cts L Rev 285, 298 (2014).

43 *Id.* at 127.