

## E-DISCOVERY | PREDICTIVE CODING

# VALIDATING THE VALIDATION SET

For predictive coding, the focus should be on the validation set, not the seed set.

BY ELIZABETH E. MCGINN, ADAM MILLER & NADAV ARIEL

**PREDICTIVE CODING IS** becoming increasingly prevalent in fulfilling discovery obligations in litigation and in response to regulatory inquiries. As the process gains acceptance, parties, regulators and courts debate whether producing parties should be required to disclose documents and coding decisions used to “train” the predictive coding software.

However, the focus on these training materials, known as the “seed set,” has shifted attention away from the more important subset of documents known as the “validation set.” The validation set, which essentially functions as an answer key, ultimately ensures the quality of the predictive coding results and should be the focus of parties, courts and regulators in determining whether a party utilizing predictive coding has satisfied its discovery obligations.



HONG LIMSTOCKPHOTO

### THE IMPORTANCE OF PREDICTIVE CODING

Predictive coding relies on an algorithm to code documents based on input received from

human reviewers. While there are various ways to implement predictive coding, the process generally involves two separate subsets of the document collection. One is

 E-DISCOVERY | PREDICTIVE CODING

the seed set, which can be created randomly from judgmental sampling or from searches designed to capture the most relevant documents. The other, the validation set, should be a statistically significant random sample of the document collection.

Reviewers manually determine whether the documents in both subsets are relevant. Based on information gleaned from the seed set documents, the software predicts whether each of the remaining documents in the overall population, including the validation set, is relevant. The accuracy of the software's predictions is then assessed by comparing its results to the manual determinations for each document in the validation set.

coding of the overall document population, including the validation set. The process repeats until the software's predictions match the manual determinations for a sufficient number of documents in the validation set. Other than the training and validation documents, none of the other documents are reviewed by humans, potentially saving significant time and money.

The software's ability to accurately predict the relevance of documents is dependent on the quality of the information it receives from the human reviewers' coding of the seed set and any additional training documents (for simplicity, we refer to both the initial seed set and subsequent training documents as the


Peck advised the attorney for the producing party that "[i]f you do predictive coding, you are going to have to give your seed set, including the seed documents marked as nonresponsive to the [opposing] counsel so they can say, well, of course you are not getting any relevant documents, you're not appropriately training the computer."

Judge Peck also indicated that the producing party's subsequent voluntary disclosure of the seed set "made it easier for the court to approve the use of predictive coding" and that the court "highly recommends that counsel in future cases be willing to at least discuss, if not agree to, such transparency."

Subsequently, disclosure of the seed set has become a major point of contention in disputes over the use of predictive coding, with a split in court opinions and discovery literature addressing the topic.

### THE FORGOTTEN SET

But all of the attention on the seed set has distracted from the more important aspect of ensuring the quality of the predictive coding process, the validation set.

 UNDETECTED ERRORS IN DETERMINING RELEVANCE IN THE VALIDATION SET CAN NEVER BE OVERCOME.

If the software's predictions do not match the manual determinations for a sufficient number of documents in the validation set, then additional documents are selected for manual review and used to further "train" the software. After this additional training, the software reassesses its

seed set). Consequently, much of the focus in the predictive coding debate has been on ensuring proper coding of the seed set.

The emphasis on the seed set was amplified by the first case approving predictive coding, *Da Silva Moore v. Publicis Groupe*, in which Magistrate Judge Andrew

## E-DISCOVERY | PREDICTIVE CODING

For the party receiving the document production, or the court overseeing the process, the most important measure of the quality of the predictive coding process is recall, which is the percentage of relevant documents in the document population that the software accurately identifies as such. Recall is measured by calculating the percentage of relevant documents in the validation set, as determined by human review, that the software correctly identifies as relevant. At the outset of the process, the recall is typically low, but it gradually improves as additional training is conducted.

Because the recall calculation depends on accurate relevance determinations in the validation set, those determinations are a critical aspect of predictive coding. Mistakes in the manual coding of the validation set will skew the recall calculation so that it may appear that the predictive coding software successfully identified a sufficient number of relevant documents when in fact it had not. Mistakes in coding the seed set, by contrast, will not have the same effect (as long as the

validation set is sufficiently sized and accurately coded) because the necessary recall will not be reached, resulting in additional training until the software is able to achieve the desired performance.

In other words, undetected errors in determining the relevance of documents in the seed set can be overcome with additional training, but undetected errors in determining the relevance of documents in the validation set can never be overcome. Therefore, parties seeking to ensure the accuracy of a predictive coding process should focus on transparency into the coding of the validation set, not the seed set.

Courts ruling on the use of predictive coding have focused on transparency into the seed set, with transparency into the validation set referenced as an afterthought, if at all.

In the recent *Rio Tinto v. Vale* case, however, Judge Peck indicated that disclosure of the seed set is not necessarily required as long as there are adequate assurances regarding recall. While the

parties in that case agreed on disclosure of all non-privileged documents in the seed and validation sets, it is possible that transparency into the validation set would have sufficed. Transparency into the validation set may also be less vulnerable to a work product claim than the seed set because it is a random sample and is not used to “train” the software to make judgments.

In sum, the emphasis on the seed set has sidetracked parties, courts, and regulators from a more important aspect of the predictive coding process, the validation set. Those using predictive coding should focus their efforts on ensuring the accuracy of the validation set above all.

---

*Elizabeth McGinn is a partner in the Washington, D.C., and New York offices of BuckleySandler LLP, Adam Miller is a partner in the Washington, D.C., office of the firm, and Nadav Ariel is an associate in the Washington, D.C., office. They advise clients on consumer financial services, white collar, e-discovery and privacy-related issues.*