



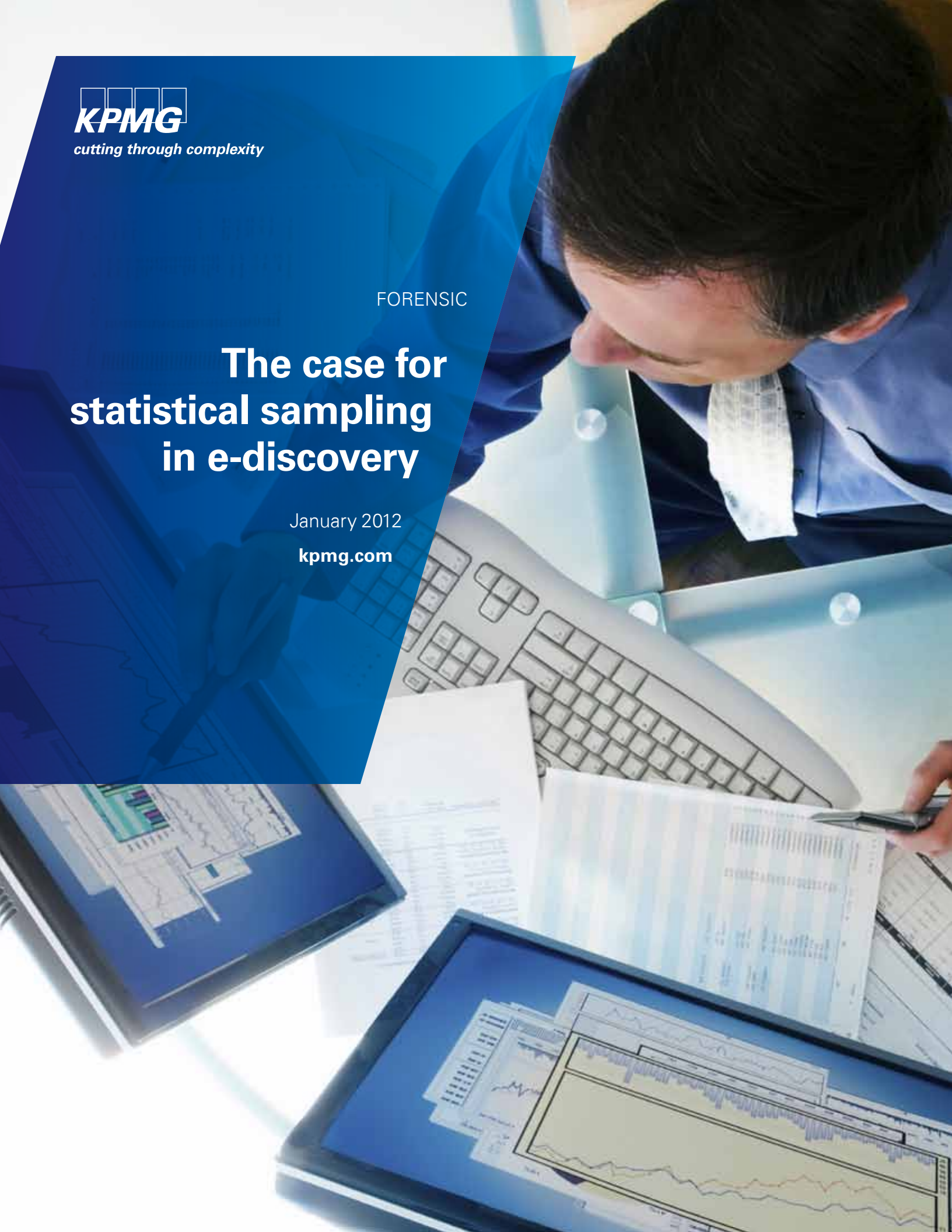
cutting through complexity

FORENSIC

The case for statistical sampling in e-discovery

January 2012

kpmg.com





The sheer volume and unrelenting production deadlines of today's electronic discovery efforts demand ever-increasing review capacity and throughput. Meanwhile, a growing chorus within the judiciary is urging adoption of technology-assisted document review. Attribute sampling in e-discovery can help litigators meet search-efficiency and quality-assurance goals, increase productivity, and reduce costs. A robust statistical sampling process can confirm whether technology-assisted results are as effective as human review or more so.

Technology and practice considerations

The increasing volume and complexity of electronically stored information and the cost of its review in connection with litigation and regulatory proceedings continue to drive the need for technology-assisted document review and the development of "predictive-coding" software. Because the majority of sophisticated high-speed processing, indexing, and categorization tools rely on sophisticated proprietary algorithms often referred to as "black-box" technologies, there has been some reluctance to use them. The fear is that the results they produce may be challenged and rejected as not meeting the required "standard of reasonableness" under the Federal Rules of Civil Procedure and other applicable rules.

Effective use of statistical sampling¹ can help overcome this concern by quantifying the reliability of a system's technology-assisted predictive-coding calls through testing and comparison of both human-reviewed and machine-assisted coding. Statistical sampling also can determine with a stated level of confidence² whether the system has produced results at least as consistent and reliable as those obtained by having attorneys review the documents manually.

Today, deciding whether an electronic discovery review process meets the reasonableness standard generally hinges on whether enough personnel have performed the review to ensure reliable results. Human (manual) review traditionally has been considered the most desirable form or "gold standard" for assessing the adequacy of a document review process.

¹ Note: This observation relates solely to process control sampling in the performance of document review in connection with electronic discovery—it is NOT a commentary on sampling techniques relative to financial statement auditing.

² *Confidence level*: The probability, expressed as a percentage, that the true value of an attribute being tested through sampling lies within a range called the "confidence interval." *Confidence interval*: An estimated range of values that likely includes the attribute value that is being tested through sampling. Example: A review of sample documents shows that 18 percent of the documents have call tag 'A.' If we are using a confidence interval of plus or minus 2 percent and a confidence level of 95 percent, then we can say that we are 95 percent confident that the actual percentage of call tag 'A' in the entire population is between 16 and 20 percent.



Results produced by manual review, however, have rarely been properly measured, questioned, or tested for consistency and reliability, and sophisticated sampling and testing of physical review processes are equally rare. Nevertheless, such review remains the dominant method despite academic and anecdotal evidence to the contrary:

The idea that exhaustive manual review is the most effective—and therefore the most defensible—approach to document review is strongly refuted. Technology-assisted review can (and does) yield more accurate results than exhaustive manual review, with much lower effort.³

In the world of electronic discovery, the myth of the superiority of human review persists, as does the fear of using statistical sampling to defend other methods of review. Rigorous statistical sampling, which is widely accepted as a basic quality-control methodology to manage and quantify quality risk in many product and service industries, has not been widely adopted in electronic discovery review projects.

Automotive operations, electronics manufacturing, and financial services, for instance, rely on sampling for quality control and setting warranty reserves, among other things; ironically, their legal counsel may eschew the same statistical processes that underpin the analysis of critical elements of their client's primary business. Scaling up traditional processes by adding more reviewers is impractical given the cost- and quality-control initiatives that are of utmost importance to corporations and the counsel on whom they rely.

Today's conditions demand new technologies

Most practitioners still review document collections, culled in some form (typically by keywords), in a linear manner. While they may organize by date, custodian, or some other parameter, every document gathered essentially starts in one pile and ends up in another—many months and dollars later. Accordingly, technology development efforts have largely focused on helping review teams manage the efficiency and cost of large-scale reviews.

Initially, tools categorized and “clustered” documents based on content or “concepts” or degree of similarity, and grouped near-duplicate documents for review. These tools prioritized to improve linear review by helping reviewers speed through nonresponsive documents or similar sets of responsive documents, bulk-tagging them to more quickly isolate material requiring further review.

Such prioritization tools also enable different documents to be reviewed at different cost levels. For instance, reviewers working on contract to law firms can weed out nonresponsive documents, allowing the firms' full-time lawyers (typically, associates) to make privilege calls and address more complex issues.

The latest technologies go beyond prioritization and clustering, and actually automate the review process by minimizing the need for human reviewers to look at every document. Senior reviewers or subject matter experts (SMEs) “train” the software to make document coding calls—predictions—based on their coding of training sets of documents. Properly configured, such software

³ “Technology-Assisted Review in E-Discovery Can Be More Effective and More Efficient Than Exhaustive Manual Review,” *Richmond J. of Law & Tech.*, Vol. XVII, Issue 3, 1-48 (2011).

review is quicker and more consistent than manual review. The results of initial training iterations can then be used to structure further review strategies.

Regardless of where an organization falls on the linear/prioritized/automated review continuum, achieving a faster, more consistent, more predictable, and less costly review takes more than just enhanced efficiency. A valid, supportable document review project requires proven quality control (QC) processes to demonstrate its effectiveness (*see sidebar*).

Cost, risk, and time: Technology vs. human review

While the cost-effectiveness of technology-based processes has been largely established, their reliability remains a topic of ongoing discussion. Over a recent three-year period, KPMG conducted four test projects⁷ that compared the results of a human review process with results obtained by reprocessing the same set of documents using a predictive-coding software tool.

The tool used in these tests is trained by an SME. The training used an initial sample of 400 documents augmented by a subsequent series of randomly selected sample batches of 40 documents each. These are reviewed consecutively by the SME until the software can reliably predict the “right” calls (those that the SME would have made).

Based on the SME’s decisions about document relevance in the training samples, the trained software calculated the relevance of the remaining documents:

In all three matters, the software-assisted review had better recall than human review; i.e., it was able to identify documents that were actually relevant that had been missed in the traditional human review workflow. In addition, the software-assisted workflow identified the subset of the total population that was likely to contain relevant documents—this subset would have to be reviewed as the next step of the workflow. In the matters we studied, the software system estimated that 93–98 percent of the relevant documents were in subsets that were only 38–68 percent of the total population.⁸

In other words, for large-scale reviews, software-assisted review consistently finds more relevant documents without the need for human reviewers to look at all of the documents. The completeness of the recall can be confirmed to a stated confidence level⁹ by statistical sampling.

⁷ Three of the test projects are described in detail in KPMG’s August 2011 white paper *Software-assisted document review: An ROI your GC can appreciate*. Similar results for a fourth project conducted in 2011 were not published.

⁸ *Ibid.*

⁹ *Ibid.* (footnote 2).

The need for proven quality control (QC) processes

In a keynote address at the July 2011 Carmel Valley e-discovery Retreat and subsequent article⁴ on the issue, United States Magistrate Judge Andrew J. Peck, United States District Court, Southern District of New York, provided support for expanded use of statistical sampling of electronic discovery review results. Statistical sampling can determine the recall and precision—the efficacy—of the predictive coding software on both a stand-alone and relative-to-human-review basis. He said:

If the use of predictive coding is challenged in a case before me, I will want to know what was done and why that produced defensible results. I may be less interested in the science behind the “black box” of the vendor’s software than in whether it produced responsive documents with reasonably high recall and high precision.

In Peck’s view, the validation of review quality is too often subordinate to the review itself, which is expressed in informal or casual observations that lack the scientific rigor and quantifiable measures necessary to defend the quality process. Judge Peck alludes to more sophisticated quality assurance methods—such as those relying on statistical sampling—to validate and definitively quantify technology-derived results.

Blogger chrisdale.wordpress.com also noted Judge Peck’s keynote endorsement of sampling and relevance ranking⁵:

If 2010 was the year of proportionality, Judge Peck said, it seems that 2011 is the year of predictive coding. He explained this as meaning (and it is his meaning which counts here, not any one provider’s definition) that a senior lawyer, using a random sample or pre-filtered set of documents decides whether they are responsive or not responsive; one goes through the process several times with senior people until the computer is sufficiently trained to apply those conclusions across a wider set or the whole. Some systems, he said, merely discriminate between relevant and non-relevant documents, while others prioritize them for review on a scale from 0 to 100. Such a system might be used to find keywords or for quality testing as well as for making primary selections. The idea is that lawyers do not spend their clients’ money reviewing irrelevant or low-rated documents.

Similarly, in *Victor Stanley v. Creative Pipe*, Judge Grimm said, “[t]he only prudent way to test the reliability of the keyword search is to perform some appropriate sampling of the documents determined to be privileged and those determined not to be in order to arrive at a comfort level that the categories are neither over-inclusive nor under-inclusive.”⁶ Such quality assurance methods as sampling need to demonstrate that technology-derived results are at least as accurate and consistent as those obtained by unaided human review.

⁴ Peck, Andrew, “Search, Forward,” *Law Technology News* (October 1, 2011) (http://www.law.com/jsp/lawtechnologynews/PubArticleLTN.jsp?id=1202516530534&Search_Forward&slreturn=1).

⁵ <http://chrisdale.wordpress.com/2011/08/02/judge-peck-and-predictive-coding-at-the-carmel-ediscovery-retreat/>.

⁶ *Victor Stanley v. Creative Pipe, Inc.* 2008 WL 2221841 (D. Md. May 29, 2008).

Measuring and assuring process quality

The ability to carefully select a sample and infer from it the condition of the larger population with a high degree of confidence in the reliability of the inference has tremendous utility in electronic discovery. In KPMG's test projects, the software consistently recalled a greater number of relevant documents than the human reviewers did.

By retrieving more of the relevant documents than a human first review and significantly reducing how much of the corpus required review, considerable savings may be achieved. The proprietary algorithm used was not disclosed, but in practice this should be increasingly irrelevant. In Judge Peck's words: "I may be less interested in the science behind the 'black box' of the vendor's software than in whether it produced responsive documents with reasonably high recall and high precision."

Demonstrating and assuring process capability to defend a technology-assisted review is a matter of (1) sound design, (2) transparency, and (3) quantifiable results. QC by process sampling is a scientifically rigorous method that delivers all three. Clearly, statistical sampling can provide benefits to QC in electronic discovery review—so why isn't it more widely used?

SOUND DESIGN

Sound design involves following an accepted QC sampling methodology, determining the appropriate confidence level, calculating the appropriate sample size, quantifying any process risks, and determining acceptance and rejection criteria.

TRANSPARENCY

Transparency involves documenting every step of the process as well as its underlying rationale, and exposing the sample and the acceptance/rejection determinations for the documents in the sample.

To quote Judge Peck once more: "That may mean allowing the requesting party to see the documents that were used to train the computer-assisted coding system. (Counsel would not be required to explain why documents were coded as responsive or non-responsive, just what the coding was.) Proof of a valid 'process', including quality control testing, also will be important."

QUANTIFIABLE RESULTS

Quantifiable results involves a meaningful discussion of the QC methodology.

Case study 1: Confidence interval sampling (Process Validation)

Statistical sampling is being successfully used in e-discovery to quantify process quality by clients who have come to understand its value in helping them meet tight time lines. One KPMG client who wanted a simple method of assessing the effectiveness of keywords requested testing on nonresponsive documents.

KPMG provided the client with alternative statistical sampling methods: (1) select a sample of 1,068 nonresponsive documents and be 95 percent confident that the sample results represent the entire nonresponsive document set; or (2) select and test 4,147 nonresponsive documents and be 99 percent confident that the results represent the entire nonresponsive document set. The entire set of nonresponsive documents was several hundred thousand.

Although these two alternatives met the client's requirements, many other sample size/confidence levels are possible, depending on the specific needs of the client.



Case study 2: Lot acceptance sampling (Process Control)

With a review under way, our client wanted to evaluate contract reviewer performance “near real time.”

After analyzing the average number of documents the reviewers were completing each day, and relying on well-established “batch” or “lot” sampling methods using a 95 percent confidence level, we provided the following sampling approach: (1) Reviewers completing 1,200 documents or fewer per day would have 80 documents randomly sampled and checked. If three or more documents were improperly coded, the entire batch, up to 1,200 documents, would be returned for rereview. (2) Reviewers completing between 1,201 and 3,200 documents per day would have 125 documents checked and, if four or more documents were improperly coded, the entire batch would be returned for rereview.

By using this batch acceptance sampling approach, the lead reviewer was able to provide timely feedback and corrective instruction to the reviewers and thus be more confident that ongoing review quality standards were being met.



Ideas for overcoming resistance

While using statistical sampling may satisfy an opposing attorney, a regulator, or the magistrate or judge presiding over the discovery proceedings, even more fundamentally, it can provide the electronic discovery practitioner with greater comfort about the accuracy of the technology used in review. So with this degree of reliability, why is there still so much reluctance to use statistical sampling?

Possible reasons include lack of familiarity with statistical (generally mathematical) methods, or concerns about mastering its perceived complexity. Another concern may be that a small error found in sampling could potentially render the entire review results unacceptable. Additionally, there seems to be broad concern that, in the discovery review process, there is a lack of legal precedent confirming the acceptability of statistical sampling methods.

The following suggestions may help move reluctant electronic discovery practitioners to overcome their bias against statistical sampling:

Note that statistical sampling is already part of your life.

Practitioners who balk at statistical sampling in electronic discovery may vigorously debate colleagues about the latest presidential poll numbers, make purchasing or investment decisions based on the latest inflation rate, or think nothing of consuming pharmaceuticals that have gone through FDA testing. It defies logic to accept statistical sampling as an integral verification tool of politics, finance, or pharmaceutical manufacturing, for instance, but reject those same methods in professional practice.

The Sedona Conference has remarked on statistical sampling in political polling,¹⁰ saying:

[Even though statistical sampling is widely used in political polling] mention “statistical sampling” in most other contexts and there is such apprehension about the process and the math involved, such that, at least to date, statistical sampling is rarely used or cited in assessing the quality and consistency of the electronic document review process. [But] the document review process is very well suited to the application of statistical sampling to improve quality and reduce costs.

Use experts. The legal profession consults experts to overcome lack of familiarity with scientific phenomena, technologies used for alleged patent infringement, and the psychology of jurors. Your company can similarly address lack of familiarity with sampling through the use of experts who understand QC sampling. Ultimately, the experts may have to explain the sampling process to an uninitiated magistrate or an overworked regulator, so why not start with your own case team?

Sampling experts also can assist with data stratification, determining sample sizes, and calculating confidence levels for statistical inferences. They can point to the generally accepted formulas on which they rely (which are widely available—e.g., in college statistics textbooks). As with all experts, cost considerations can be evaluated using a cost-benefit calculation that compares the cost of the experts to the avoided costs of more extensive testing using nonstatistical approaches.

¹⁰ See “Sampling 101 for the E-Discovery Lawyer,” an appendix to the May 2009 Sedona Conference Commentary on Achieving Quality in the E-Discovery Process.

Take incremental steps. Another approach to adopting the use of statistical sampling would be to programmatically integrate generally accepted QC sampling methods into widely used electronic discovery applications. One does not have to use predictive coding to benefit from sampling in electronic discovery. Keyword-searching, which has become mainstream, relies on somewhat-arcane search techniques and could benefit from statistical sampling and validation.

Numerous examples of statistical sampling in electronic discovery (other than predictive coding) exist. Applying a new sampling methodology could include periodic batch sampling throughout the review process with a mechanism designed to address reviewer error as it is detected, thus reducing the need to repeat a significant portion of the review process. Others include testing potential keywords (inferring responsiveness rates based on the existence of a particular keyword), comparing responsiveness rates for different types of documents (e-mail vs. loose files), or other examples described in *The Sedona Conference Commentary*.¹¹

Carefully designed user interfaces for selecting samples, testing them, and reporting the results can guide users through the sampling process, minimizing if not eliminating most common sampling mistakes. This would increase consistency, repeatability, and reproducibility in the QC process, and would help increase familiarity with statistical methods and their defensibility in electronic discovery.

Sample your sampling. “Sampling error” caused by errors in evaluating the sample items can lead to invalid interpretation of the sample results. This can cause extra work to rereview documents or significantly expand the sample size. The risk of sampling error can be addressed with a sample evaluation process in which senior SMEs check the sample results before they are finalized.

Conclusion: Hurry up and wait (but not for long)

The lack of a clear legal precedent is likely to be addressed in the near future by the courts, which are becoming increasingly aware of the benefits of new technology and statistical sampling in dealing with the challenges posed by very large populations of documents. Even without a definitive Federal District Court opinion on the subject, members of the bench are trying to speed this process. “[U]ntil there is a judicial opinion approving (or even critiquing) the use of predictive coding, counsel will just have to rely on this article as a sign of judicial approval,” Judge Peck has said.

Without clear legal precedent, there may be some additional risk to applying new technologies and relying on statistical sampling to demonstrate their efficacy. Companies will have to decide whether the benefits in terms of quality and cost of QC sampling that result from these new technologies can more than offset such risks until the legal precedents supporting their use are

clearly established. Until then, even the most reluctant electronic discovery practitioner can begin to familiarize him- or herself with statistical methods to get more comfortable with incorporating them into their existing electronic discovery processes.

About the authors

Chris Paskach is the national leader for KPMG’s 220-member Forensic Technology Services practice. He coordinates the full range of KPMG’s eDiscovery services, from records and information management, digital evidence recovery, and evidence and discovery management services, including KPMG’s secure eDiscovery hosted review repository services. As a member of *The Sedona Conference Working Group 1*, he participated in the drafting of *The Sedona Conference Commentary on Achieving Quality in the E-Discovery Process* (May 2009) and *The Sedona Conference Database Principles* (March 2011).

Michael Carter is a director in KPMG’s eDiscovery practice, leading its process improvement deployment. His responsibilities include the technical integration of statistical sampling into KPMG’s eDiscovery processes and technology. Michael is a certified Six Sigma Master Black Belt and member of the American Society of Quality (ASQ).

Phil Strauss is a director in the firm’s San Francisco office, coordinating KPMG’s eDiscovery and other services for Northern California and the Pacific Northwest. He has experience in discovery and litigation matters from private practice with AmLaw firms, as general counsel of two public software companies, and as an outside consultant. He is on the adjunct faculty of the University of California, Berkeley’s Haas School of Business, where he teaches business law.

¹¹ Ibid.

For more information, please contact:

Christopher H. Paskach, CPA

Partner

T: 714-934-5442

E: cpaskach@kpmg.com

Michael J. Carter

Director

T: 714-934-5496

E: mjcarter@kpmg.com

Philip R. Strauss, JD

Director

T: 415-963-5156

E: philipstrauss@kpmg.com

kpmg.com

This article represents the views of the authors only, and does not necessarily represent the views or professional advice of KPMG LLP.

The information contained herein is of a general nature and is not intended to address the circumstances of any particular individual or entity. Although we endeavor to provide accurate and timely information, there can be no guarantee that such information is accurate as of the date it is received or that it will continue to be accurate in the future. No one should act upon such information without appropriate professional advice after a thorough examination of the particular situation.

© 2012 KPMG LLP, a Delaware limited liability partnership and the U.S. member firm of the KPMG network of independent member firms affiliated with KPMG International Cooperative ("KPMG International"), a Swiss entity. All rights reserved. Printed in the U.S.A. 25582NSS

The KPMG name, logo, and "cutting through complexity" are registered trademarks or trademarks of KPMG International.