AHLA 50 1967-2017

January 2017 Volume 21 Issue 1

connections

For the health and life sciences law community

Health Law Issues 2017

(page 27)

OFFICIAL MAGAZINE OF AMERICAN HEALTH LAWYERS ASSOCIATION

Using and Defending Against Statistical Sampling in False Claims Act Cases

By Jonathan Phillips and Julia Reese, Gibson Dunn & Crutcher LLP, and JoAnna Younts, Berkeley Research Group

In recent years, the number of health care fraud allegations under the False Claims Act (FCA) targeting large numbers of claims for reimbursement from government health programs has seemingly proliferated. That is because the government and FCA whistleblowers (also known as "relators") often develop their allegations to cover the broadest possible universe of claims related to a single theory, such as claims by all providers in a single practice, or by all health care facilities under common ownership or management.

But in many of these cases, the amount of specific, direct evidence that FCA plaintiffs are adducing has not necessarily increased in line with the breadth of the claims at issue. The government and relators are increasingly turning to statistical sampling and extrapolation to determine the extent of alleged FCA violations. Specifically, this entails reviewing just a small sample of the target providers' claims to determine which, if any, were not compensable and therefore constituted "false claims." FCA plaintiffs then use statistical methods to extrapolate the percentage of false claims in that sample to the full number of the providers' claims during the relevant time period to determine the overall FCA exposure.

Statistical sampling as a means to expand the universe of actionable claims becomes particularly notable considering that the FCA imposes sanctions of treble damages and potential civil penalties of up to $22,000 per claim. So it is unsurprising that the use of statistical sampling has drawn increased attention from litigants and the courts. While some FCA defendants have argued that statistical sampling is legally improper, recent court decisions have looked at the role of statistical sampling as an evidentiary issue, not a per se legal issue. But even as a number of courts have allowed statistical sampling as a form of evidence, they also have provided useful commentary that may guide how parties approach statistical sampling in their cases.

Statistical sampling is a potentially important tool outside of litigation as well, and is certainly not limited to the government's use. Providers often find sampling to be a cost-effective tool for developing their defenses, understanding their litigation risk, and supporting pre-suit discussions with the government. In any context, the sampling approach is susceptible to abuse, manipulation, or simply erroneous deployment of statistical methodologies. By making sure samples are tailored to the specific facts and legal theories of the case at hand, and employing sound methodologies, providers can better examine the potential exposure and develop better responses—and possible legal challenges—to FCA plaintiffs' attempts to expand liability through sampling. This article highlights some of the recent discussion in the courts about the evidentiary use of statistical sampling that providers should be thinking about in their FCA cases. It then details some of the most important considerations behind developing a sound approach to sampling to ensure that it is done effectively.

Legal Landscape

The use of statistical sampling in FCA cases is not necessarily new. Earlier cases, though, tended to involve the use of statistical methods to determine the number of affected claims where the alleged wrongdoing either was not contested or otherwise already established. That was consistent with the distinction, drawn over 80 years ago in the seminal antitrust case *Story Parchment Co. v. Paterson Parchment Paper Co.*, between evidence presented for the purpose of establishing the *extent* of liability where the damages are "definitely attributable to the wrong," and evidence introduced to establish the *fact* of liability where the connection between the wrong and the alleged injury is in dispute.[1] Many courts have long been willing to accept statistical sampling under the former circumstances (although not without exception), often in the name of the efficient deployment of judicial resources where a claim-by-claim review would be impractical.[2]

More recently, though, the government and relators have relied on statistical techniques for the latter purpose: to establish that certain claims are actually fraudulent without ever directly examining them, instead assuming the claims are violative based on extrapolation from medical review of a small sample of similar claims. In other words, these plaintiffs look to sampling to establish that a claim was "false" or "fraudulent" under the FCA without ever looking at it.

The Supreme Court seemingly endorsed this method of proof, at least in some circumstances, in *Tyson Foods Inc. v. Bouaphakeo*. In *Tyson Foods*—a class action involving allegations that an employer systematically undercompensated employees for time spent donning and doffing equipment—the employees introduced an analysis of a supposedly representative sample designed to establish the extent to which each employee in the certified class was undercompensated.[3] (Individualized evidence was not available, in part because the employer did not keep track of the uncompensated time.) The employer objected to the use of representative samples to meet classwide elements essential to liability, but did not offer evidence specifically designed to undermine the statistical methodologies employed by the plaintiffs' expert.[4]

The Supreme Court declined to adopt a per se rule, holding instead that "[w]hether and when statistical evidence can be used to establish classwide liability will depend on the purpose for which the evidence is being introduced and on the elements of the underlying cause of action."[5] While no federal circuit court has yet squarely addressed this issue in an FCA case, trial courts also have resisted calls for categorical prohibitions on statistical sampling to establish FCA liability—or at least the "falsity" element.[6]

These courts have instructed that statistical sampling is but one form of evidence that plaintiffs can use to establish elements of FCA liability. As such, even in cases in which the court refuses a categorical prohibition, statistical sampling is subject to attack, and potentially exclusion under the rules of evidence, for lack of validity and reliability. Indeed, as one court stated in *United States ex rel. Ruckh v. Genoa Healthcare*, although "no universal ban on expert testimony based on statistical sampling applies in a qui tam action," "defects in method, among other evidentiary defects, might result in exclusion."[7]

The decisions in *Tyson Foods* and *Genoa Healthcare* highlight two key considerations for guarding against the inappropriate use of statistical evidence during the liability phase of an FCA case. First, although courts may be reluctant to establish a per se prohibition, the theories in the case at hand must be susceptible to proof by extrapolation. Second, even where extrapolation may be permissible in theory, the methodology employed by the plaintiffs must be statistically valid and, once again, tailored to the specific facts and circumstances of the case at hand.

Another recent FCA decision, *United States ex rel. Wall v. Vista Hospice Care, Inc.*, illustrates both of these considerations. First, in *Vista*, the court rejected the reliability of extrapolation to determine liability for hospice claims, because the eligibility determinations at issue by their nature require the careful use of subjective clinical judgment about each individual patient's life expectancy.[8] In other words, because of the nature of the claim at issue, extrapolation could not possibly serve as a substitute for direct proof on a claim-by-claim basis. The court rejected the notion that judicial (and party) economies should dictate a different result—and indeed, even rejected the idea that a claim-by-claim review of all 12,000 claims at issue would be impractical—noting that the relator's decision to pursue liability for a large number of claims did not have the effect of lowering her burden of proof.

Second, the *Vista* court alternatively held that the relator's statistical evidence was inadmissible because the "methodology was fundamentally flawed."[9] Specifically, the court found problematic the unexplained exclusion of certain claims and duplication of others in the universe from which the sample was drawn, and the stratification of the sample along lines that turned out to be inconsistent with the relator's theory of liability. Beyond that, the expert's sampling design failed to control for certain relevant variables, including differences in eligibility determinations that might be attributable to each patient's illness and/or different clinical practices among different physicians in different geographic regions. As a result of these "fatal" flaws, the court found the expert's extrapolation "unreliable, even if it is assumed to be generally allowable."[10]

The court's analysis in *Vista* shows that FCA litigants can and should focus attention on the soundness of the proposed statistical methodologies and the nexus between those methodologies and the proponent's theories. Indeed, it may be the case that the specific theories at issue are not susceptible to proof by extrapolation at all. But even in cases in which courts are open to the use of sampling evidence as a general matter, the reliability and probative value of that evidence can and should be closely evaluated for potential challenge under the rules of evidence. The next section outlines some of the factors to consider when evaluating the propriety of statistical methodologies to ensure that the analytical approach is consistent not only with generally accepted statistical principles but also the theories forwarded in a particular case.

Statistical Methodology Considerations

As noted above, relators and the government find many FCA investigations to be well-suited for sampling because they typically involve large numbers of claims of the same type where the same billing or processing activities are repeated (and because that process is much less costly, and potentially much more rewarding, than a claim-by-claim analysis). However, despite these unifying characteristics that might suggest an FCA case is a good candidate for proof by statistical evaluation, there is a need to select samples to fit each situation rather than using a "one-size-fits-all" approach.

While the selection of the provider or supplier and the selection of the time period are relatively straightforward, one of the most important issues in sampling and extrapolation is determining the questions that need to be answered, or what specifically is expected to be learned from the data—in legal parlance, the theory or theories the analysis is supposed to support. This helps to inform not only the definition of the "sampling universe," but also the manner in which the sample should be drawn.

Defining the Sampling Universe

As suggested above, the proper sampling universe may vary depending on the type of medical service(s) represented in the claims, the provider type(s) that submitted the claims, the time frames when the services were delivered and/or when the

claims were submitted, and other factors. Subsets of claims to be included in or excluded from the universe should be identified, along with the reason for those decisions.

For example, if the question at issue relates to the medical necessity of imaging studies billed during a specific four-year period, it must be determined whether the universe constitutes all radiology-related CPT codes, or just CT and MRI codes, or some other subset. In another case, the question may be whether there was medical necessity to support the number of patients admitted to the hospital by a specific provider. In this case, it would be important to determine whether the universe should consist of all inpatient claims or just those for certain Diagnosis Related Groups (DRGs).

Developing a Sound Sampling Plan

Once the universe of claims is well defined, a sampling plan can be developed. Generally, the "sampling unit" is a single claim,[11] and the "sampling frame" is the list of all the possible sampling units from which the sample is selected.

Similar to selecting the universe, the importance of asking the right question(s) comes into play again in designing the sampling plan, and will help determine whether a simple random sample would be appropriate or whether a more nuanced analysis is necessary. As noted above, mistakes in sample selection can introduce errors that will wholly invalidate the statistical analysis or undermine the nexus between it and the legal theories. Issues that should be considered include the following:

- Are there differences in types of patients, treatments, or providers that need to be addressed in the sample?
- Are there changes or differences in treatment protocols, practice guidelines, or standards of care over the time period in question?
- If multiple time periods and/or geographies are relevant, are there differences and changes in government regulations, national or local coverage determinations, or reimbursement methods that need to be considered?

In other words, the theory of liability should be closely evaluated to determine the factors that should inform the sampling design.

In many FCA cases, the theory of liability/damages may lend itself to a comparison of claims across different categories. In these cases, a "cluster" or "stratified" sample may be appropriate. If the universe is relatively heterogeneous and the claims naturally "cluster" into the different categories, then the claims may simply be sampled from these natural clusters.[12] If the claims are not heterogeneous and don't naturally categorize themselves, then an expert may segment the claims into different groups using some variables consistent with the legal theories, and select an appropriate sample from each resulting strata.[13] While cluster samples and stratified samples facilitate similar analyses, the former increases sampling errors while the latter reduces them.[14]

Unlike a simple random sample, in a stratified sample, it is not necessary for each claim in the universe to have an equal chance of being selected. Instead, the probability of selection must only be equal within the strata. Stratified random sampling has two basic requirements: (1) the strata must be "mutually exclusive," i.e., no single item can be included in more than one group, and "exhaustive," i.e., every item must be included in one of the groups; and (2) the sample taken within each strata must be a simple random sample.[15]

Finally, it is important to understand that "random" is not the same as "representative" when it comes to samples. If a sample is representative of a population, statistics calculated from sample data will be close to corresponding values from the population. Therefore, representativeness should be tested if the sample has been selected by another party such as the Department of Justice (DOJ), the Department of Health and Human Services (HHS) Office of Inspector General (OIG), a relator, or an auditor.

Use of RAT-STATS to Select Samples

OIG developed RAT-STATS in the 1970s for a variety of purposes, including self-assessment audits as part of internal compliance programs and evaluating provider self-disclosures to OIG and the Centers for Medicare & Medicaid Services (CMS). But RAT-STATS gained substantial traction with the creation of the Health Care Fraud Prevention and Enforcement Action Team (HEAT) in 2009 by HHS and the DOJ, as combatting Medicare fraud became a federal priority. Its functionality in generating random samples and estimation has allowed its increasing use in sampling and extrapolation in numerous FCA investigations.

RAT-STATS has some important core functions, including sample selection. It also contains estimation modules that allow for various kinds of extrapolation, such as "variable" and "attribute" appraisal. Variable appraisal is typically used for overpayment estimation. "Attribute" appraisal is used to estimate occurrences or percentages, such as accuracy rates, participation rates, or error rates. The CMS Medicare Program Integrity Manual Chapter 3 describes the specific standards that must be complied with when sampling and extrapolation are used for overpayment purposes.[16]

Attorneys and experts often debate the use of RAT-STATS over other alternative statistical software. One of the primary advantages of using RAT-STATS is that it is free. Although it does not have the sophisticated user interface that other software packages have, it is generally user-friendly. In addition, it is sometimes easier to simply use the same software to validate and analyze the government's chosen data in investigations than to spend valuable time and money arguing the merits of an alternative. However, it is important to understand that there are, in fact, alternatives, such as SAS, Stata, and SPSS, each of which have potential advantages over RAT-STATS depending on the situation and user preference.

Understanding Error Rates

An error rate, in its simplest sense, is an occurrence rate. However, in the context of claims auditing or investigations, the error rate generally is considered to be the improper payment rate. For example, CMS' Comprehensive Error Rate Testing (CERT) program measures the error rate in Medicare fee-for-service claims by reviewing a stratified random sample of submitted claims to determine whether the claims met Medicare requirements. Overpayments and underpayments are then calculated based on the error rate. CMS' target error rate in the CERT program is 5.4%.[17] The Medicaid program has a similar auditing program called Payment Error Rate

Measurement Program (PERM). OIG Corporate Integrity Agreements (CIA) also often incorporate heightened review of submitted claims if providers subject to a CIA reach a threshold error rate.

All of these programs define error rates in terms of dollars. However, for internal purposes, it is also important to distinguish between the claim error rate and the financial error rate. The claim error rate is the percentage of *claims* that do not meet the payer requirements while the financial error rate is the percentage of *dollars* paid incorrectly, and these can be very different. A small number of incorrect claims (low claim error rate) can produce a high dollar amount of incorrect payments (high payment error rate).

In evaluating FCA liability, the definition of the universe and design of the sample are critical in laying the groundwork for the analysis, and then calculating and understanding the error rates shed light on the scope and scale of the issue. But error rates will naturally vary across organizations and types of services, and may even vary inherently depending on what the allegation or issue is.[18] Indeed, some level of error in claims of any kind is not surprising, and as demonstrated in Medicare's CERT program, CMS establishes an aggregate target error rate for the program but accepts variation across different types of services. CMS' Medicare Program Integrity Manual similarly describes different ways for Medicare Administrative Contractors to address different error rates.[19]

Likewise, in investigations of fraud and other non-contractor investigations, the level of error rates that is considered acceptable can naturally vary further still. In some cases, an error rate is deemed "too high" or "unacceptable" without any context regarding how high or how low would be considered acceptable for the potential legal claims being investigated. Therefore, it is important to understand how the errors are being defined and how the error rate is being calculated. In addition, providers would be wise not only to understand how the government is performing calculations, but also to consider statistical sampling as a defensive tool. Providers in these situations should independently identify the universe of claims they believe to be at issue and the rationale, as well as design a sampling plan using accepted statistical sampling techniques as described above. This may help identify potentially problematic statistical methods early in an investigation.

Conclusion

While CMS and third-party contractors have long used statistical sampling as a means of identifying potential overpayment issues, relators and the DOJ have begun to employ statistical methods more frequently not only to identify the scope of fraud damages, but also to establish the elements of fraud in the first instance. This significantly raises the stakes for providers, and underscores the extent to which a sophisticated, nuanced understanding of sampling design is essential for any adequate defense of health care fraud allegations. By understanding the critical elements of appropriate statistical methods and how they should be tailored to the specifics of the case at hand, defendants can better respond to, and defend against, the use—and perhaps misuse—of sampling and extrapolation.

About the Authors

Jonathan M. Phillips is a Senior Associate in Gibson Dunn's FDA and Health Care practice group and a former Trial Attorney in the U.S. Department of Justice's Civil Fraud section. His practice includes counseling health care and life sciences clients in a range of government investigations and enforcement actions.

Julia L. Reese is an Associate in the Palo Alto office of Gibson Dunn. She practices in the firm's litigation department, and is a member of the FDA and Health Care and White Collar Defense and Investigations practice groups.

JoAnna Younts is a Director in BRG's Healthcare practice. She provides strategic advisory services regarding coding, billing, and reimbursement to providers, health plans, and other health care organizations.

Endnotes

1 *Story Parchment Co. v. Paterson Parchment Paper Co.*, 282 U.S. 555, 562 (1931).
2 *See, e.g., United States v. Fadul, No.* CIV.A. DKC 11-0385, 2013 WL 781614, at *14 (D. Md. Feb. 28, 2013) (accepting the government's extrapolated damages estimate in an FCA case in which the defendant conceded liability, in part because "a claim-by-claim review [was] not practical").
3 *Tyson Foods, Inc. v. Bouaphakeo,* 136 S. Ct. 1036, 1042–45 (2016).
4 *Id.* at 1044, 1048–49.
5 Id. at 1046.
6 *See, e.g., United States ex rel. Ruckh v. Genoa Healthcare,* LLC, No. 8:11-CV-1303-T-23TBM, 2015 WL 1926417, at *4 (M.D. Fla. Apr. 28, 2015); *United States v. Robinson, No.* 13-CV-27-GFVT, 2015 WL 1479396, at *11 (E.D. Ky. Mar. 31, 2015) (holding that the defendant had "not shown that the government's use of statistical sampling in this matter or extrapolation of damages is improper as a matter of law").
7 *Ruckh*, 2015 WL 1926417, at *4.
8 *United States ex rel. Wall v. Vista Hospice Care*, Inc., No. 3:07-CV-00604-M, 2016 WL 3449833, at *12 (N.D. Tex. June 20, 2016).
9 *Id.* at *13.
10 *Id.* at *13–14.
11 It is important to distinguish between a claim, which includes all services billed for a specific office visit, inpatient admission, or other medical encounter, and claim lines, which are the individual items, such as services identified by CPT codes, included in a claim.
12 *See, e.g., Ctrs. for Medicare & Medicaid Servs., Medicare Program Integrity Manual* chs. 8.4.4.1.4, 8.4.11.2 (2016), *available at* https://www.cms.gov/regulations-and-guidance/guidance/manuals/internet-only-manuals-ioms-items/cms019033.html.
13 *See, e.g., id.* chs. 8.4.4.1.3, 8.4.11.1.
14 *See, e.g., id.* ch. 8.4.11.2 ("Selecting payments in clusters rather than individually usually leads to a reduction in the precision of estimation.").
15 *See, e.g., id.* ch 8.4.4.1.3.
16 *Id.* ch. 3.5.2.
17 *U.S. Dep't of Health & Human Servs., Office of Inspector Gen.,* OEI-09-12-00090, *Medicare Claims Administration Contractors' Error Rate Reduction Plans 2* (2015), *available at* https://oig.hhs.gov/oei/reports/oei-09-12-00090.pdf.
18 *See, e.g., Ctrs. for Medicare & Medicaid Servs., Medicare Fee-for-Service 2012 Improper Payments Report 5*, 36 (2012), *available at* https://www.cms.gov/Research-Statistics-Data-and-Systems/Monitoring-Programs/CERT/Downloads/MedicareFeeforService2012ImproperPaymentsReport.pdf (noting varying error rates for different service types such as inpatient hospital claims, DMEPOS, and physician services).
19 Medicare Program Integrity Manual, *supra* note 12, at ch. 3.7.1.2.