Coding Case Law for Public Health Law Evaluation

A Methods Monograph

PHLR

Making the Case for Laws that Improve Health

PUBLIC HEALTH LAW RESEARCH

November 26, 2011

Robert Wood Johnson Foundation

Coding Case Law for Public Health Law Evaluation

A Methods Monograph
for the
Public Health Law Research Program (PHLR)
Temple University Beasley School of Law

By:

Mark A. Hall, J.D.
Professor of Law and Public Policy
School of Law
Wake Forest University

PHLR is a national program of the Robert Wood Johnson Foundation

Summary

This monograph explores the special considerations in coding text when the relevant legal materials are judicial decisions. The content of case law merits careful study not simply because judicial opinions reflect or respond to the law, but because they are the law. But, more than this, judicial opinions are detailed repositories that show what kinds of disputes come before courts, how the parties frame their disputes, and how judges reason to their conclusions.

Content analysis of case law boils down to three steps: 1) selecting cases; 2) coding cases; and 3) analyzing (often through statistics) the case coding. Insights gained from content analysis of large numbers of opinions supplement the deeper understanding of individual opinions that comes from traditional interpretive legal research techniques. The content of judicial opinions can be important in the study of the broader social, economic, and political systems that interact with judicial precedent, but cases are also well worth scientific study in their own right. For instance, content analysis can identify previously-unnoticed patterns that warrant deeper study, or sometimes correct misimpressions based on more ad hoc surveys of atypical cases.

The major limitation of content analysis is that facts and reasons recorded in judicial opinions cannot be treated as accurate and complete. Therefore, researchers should be cautious about the meanings they attach to observations made through content analysis. With this caveat in mind, I describe a range of acceptable and best practices for systematically selecting relevant cases, forming coding categories, training coders, and testing for coding reliability.

Introduction

This monograph explains how the research method of content analysis can be applied to study legal decisions. Content analysis is a method for systematically reading and analyzing "texts" of any kind. Developed by sociologists and political scientists, the method is also used widely in the communications field (Krippendorff, 2004; Neuendorf, 2007). It can be applied not only to conventional written material, but also to images or audio content. Written texts abound, of course, in the legal sphere, including statutes, regulations, hearing transcripts and court filings.   But our sole focus here is one especially important legal text: judicial opinions.

The content of judicial opinions merit careful study not simply because they reflect or respond to the law, but because they are the law. Legal researchers are correct to recognize that it "is almost impossible to study law in a meaningful way without some attention to the [content of] opinions that contain these justifications" (Friedman, 2006, p. 266). For instance, a researcher wanting to know what effect the First Amendment's protection of speech, religion and association might have on enforcement of various public health laws would need to analyze Supreme Court opinions as a primary source of law.

But, more than this, judicial opinions are detailed repositories that show what kinds of disputes come before courts, how the parties frame their disputes, and how judges reason to their conclusions. Thus, for instance, Harr and colleagues (1977) coded for the presence or absence of 167 different factors in each of 79 zoning dispute cases decided by one state's Supreme Court over a 25-year period, to determine which factors appear to influence the outcome of these cases. It is this factual and analytical richness of judicial opinions that establish both their substantive legal importance and their utility as instruments for public health research of various designs.

On the surface, content analysis appears simple, even trivial, to some. It boils down to three steps: 1) selecting cases; 2) coding cases; and 3) analyzing (often through statistics) the case coding. The method comes naturally to legal scholars because it resembles the classic scholarly routine of reading a collection of cases, finding common threads that link the opinions, and commenting on their significance (Hall & Wright, 2008). But content analysis is much more than a better way to read cases. It brings scientific rigor to the collection and analyses of case law, which could create a distinctively legal form of scientific empiricism. This approach to reading cases can be used profitably in three distinct types of studies: those that (1) identify determinants of judicial decision-making, (2) measure consequences of judicial decisions, or (3) document how the judicial system operates.

What Content Analysis Can and Cannot Tell Us

Content analysis is not a panacea. It has certain advantages, along with substantial limitations, compared with conventional legal analysis. At best, the method generates objective, falsifiable and reproducible knowledge about what selected courts do and how and why they do it. But, this method does not lead to the Holy Grail of a true legal science. It works best when each of the judicial opinions in a collection hold essentially equal value, but not where what is needed is a deeply reflective understanding of a single pivotal case. Content analysis, therefore, does not displace traditional interpretive legal scholarship. Nor does it reveal how all aspects of a legal system function (Hoffman, Izenman, & Lidicker, 2007). Instead, it offers distinctive insights that complement the types of understanding that traditional legal analysis can generate, or that could be obtained by more direct observation of legal systems.

Conventional Legal Analysis

Traditional legal scholarship relies, like the interpretation of literature, on the interpreter's authoritative expertise to select important cases and to draw out noteworthy themes and potential social effects that might flow from the decisions.   The audience depends on the author's judgment about which cases are worth reading, which are the "leading cases" that best illustrate the historical moment in question. Interpretive legal scholars read opinions closely, looking for common themes running through several opinions. They ponder the meaning of a decision for future cases by asking how the outcome in the current case relates to its facts, procedural posture, and the court's reasoning.

Although legal writing in this mode may contain many assertions about how judges think or act, it is not a scientific form of empiricism. These legal analysts report what they see in key cases and how they interpret these observations, not unlike a literary critic might interpret poetry. Establishing some plausible basis for the minimally empirical claims in such work is usually done simply by citing relevant sources that readers can verify if they wish.

Although content analysis has different epistemological aims, it can be seen as a logical extension of the school of jurisprudence known as Legal Realism.   Over a century ago, Justice Oliver Wendell Holmes, Jr. famously proclaimed that "prophecies of what the courts will do in fact, and nothing more pretentious, are what I mean by the law." This credo, once revolutionary, is now so widely accepted that it is sometimes said in the legal academy that that "we are all legal realists." All that content analysis seeks to do is to use accepted scientific methods to support the verifiable claims that legal researchers frequently make about what judges do and say.

Content analysis can augment conventional analysis by identifying previously-unnoticed patterns that warrant deeper study, or sometimes correcting misimpressions based on more ad hoc surveys of atypical cases. Once detected, these previously unnoticed and unexpected features of the law, observed only on the surface, can be explored more deeply through other, richer methods. Scientists speak in terms of "triangulating" different methods – that is, exploring whether different approaches offer similar conclusions, each approach rigorous in its own way, but each illuminating different dimensions and potentially overcoming each others' respective shortcomings. Quantitative description can tell us the what of case law; other methods may be better suited to understanding the why and wherefore.

Neither type of scholarship standing alone is as strong as the different types combined. Content analysis reaches a thinner understanding of the law than that gained through more subjective interpretive methods. The coding of case content does not fully capture the strength of a particular judge's rhetoric, the level of generality used to describe the issue, and many other subtle clues about the precedential value of the opinion. Or to put the point more bluntly, the "legal and cultural salience of Roe v. Wade far outruns is statistical significance" (Goldsmith & Vermeule, 2002).

Inevitably, then, content analysis trades depth for breadth. Content analysis is valid if it accurately and reliably measures the particular components of the decision that the researcher wants to study. Using systematic defined coding protocols improves measurement by removing elements of researcher bias and improving thoroughness, precision and accuracy. However, to the extent that content analysis cannot reach important aspects of legal interpretation that are impossible to code objectively (such as nuance related to infrequent or highly complex factual and procedural patterns), plain content analysis loses relevance. When uniform coding cannot capture important details about idiosyncratic decisions, content analysis alone is not capable of measuring what lawyers or scholars would consider to be a full and accurate statement of the law.

Still, content analysis holds value not only for conventional doctrinal analysis, but also for more theoretically-influenced work in major branches of jurisprudence, such as economic analysis or critical theory. Writers in each of these scholarly camps frequently claim, for instance, that judges and law respond predictably to various social, political and market conditions—empirical claims that can be tested systematically. Some schools of jurisprudence emphasize the bare outcomes of cases in relation to their raw facts, while others emphasize the importance of how judges explain their decisions. Although these fundamental differences hugely affect how content analysis might be employed, the method itself is adaptable to any branch of legal theory that systematically studies judicial opinions. Moreover, the method is a key component of empirical studies evaluating the effects of case law on the public's health. Such evaluative studies require transparent and reproducible numeric indices of law and how it varies across time and across jurisdictions, to examine how such variations affect population-level health-relevant exposures, behaviors and outcomes.

Counting Case Outcomes

One basic use of content analysis is simply to document the bare outcomes of cases. Measuring who won and who lost differs fundamentally from measuring the law of a case. Case outcomes are much narrower and more objective questions, requiring much less legal judgment, than what legal principle a case embodies. One might, for instance, want to simply tote up the number of cases, and who won and lost, in which the authority of local public health laws was challenged, either in one jurisdiction or many, over a period of time.

Counting case outcomes in this fashion is best done when each decision should receive equal weight, that is, when it is appropriate to regard the content of opinions as generic data. Coding and counting cases usually assumes that the information from one opinion is potentially as relevant as that from any other opinion. Because content analysis tends to regard all cases, judges, courts, and jurisdictions the same, it should be used only with great caution when any of these have a great deal more status or influence than the others, for the question addressed.   Differential influence is often true in legal analysis because precedent and persuasiveness depend on various qualitative judgments about the reasons given or the source of the decision.

Taking this limitation into account, scholars have found that it is especially useful to code and count cases in studies document the absence of some element that is thought to be present in case law. "Proving a negative" is much harder than simply pointing to what is present in case law because the nay-saying researcher needs to demonstrate that he or she has looked exhaustively for all likely instances of the missing element.

Counting cases can also be useful in studying a wide range of social and economic phenomena that might affect judge-made law. Treating case outcomes as the dependent variable, the range of potential influences on judicial behavior that might be studied statistically is limited only by the bounds of a researcher's imagination. One study, for instance, explored whether the political make-up of Congress or changes in the Presidential party in power affected the outcome of federal appellate cases where health and safety regulations were challenged, over a 25 year period (Revesz, 1997; Revesz, 2001).

Case law can also be used as an independent variable, by asking how it influences various social and economic conditions. Law's effect on society is obviously a rich field of inquiry, but most such studies trace the effects only of statutory or regulatory law. Researchers have neglected the possible effects of judge-made law, including statutory interpretation. For instance, social host and bar owner liability for alcohol-related injuries is determined by both judicial decisions and statutory enactments, which vary widely by state and over time. These differences might contribute to a variety of public health effects of interest. With its diverse "laboratory of states," the U.S. offers boundless opportunities to learn from the "natural experiments" created by the inevitable differences in case law among jurisdictions and over time.

Evaluating Legal Doctrine

Opinion-coding is not suited, however, to evaluating the legal correctness of judicial opinions. Certainly, many content analysts draw normative implications from what they observe, but their coding of cases aims only to document what judges do rather than to evaluate in a formal empirical manner how well they perform. Without an independent "gold standard" for what the law should be in any particular case or jurisdiction, who is to say its judges are legally wrong, in an empirical sense? After all, what judges say is the law. Therefore, normative evaluation of legal doctrine ordinarily can be done convincingly only through some form of traditional legal analysis.

But, beyond documenting merely the bare outcomes of legal disputes, content analysis might be used to study the legal principles one can extrapolate from those outcomes and the facts and reasons that contribute to those outcomes and principles. Such analyses raise important epistemological and jurisprudential issues.

"Jurimetrics." The most ambitious use of content analysis is to study the legal factors that determine the outcomes of cases, using sophisticated statistical methods to model or predict the behavior of judges. This general approach at one time was called "jurimetrics" (Loevinger, 1961). This kind of study might, for instance, attempt to predict the likely result in a case when the parties present the judge with a particular combination of legally relevant factors. Often, the stated purpose is to help practicing lawyers make better-informed decisions about handling particular cases. Other times, the purpose is more scholarly—to test various claims based on legal theory.

An especially interesting subgenre uses content analysis to find some order and logic in a body of case law that, by conventional analysis, appears chaotic or haphazard. As Fred McChesney (1993) notes, "the academic history of American law generally is replete with instances in which scholars have proclaimed traditional common-law modes of distilling 'the law' from cases unworkable." These conventional legal analysts, throwing up their hands, conclude that the law on the topic is hopelessly confused and inconsistent, or less pejoratively, dependent on individual facts. Nuisance law might be one relevant example. Are public nuisances solely "in the eye of the beholder," or are there patterns of factors that are associated with the likelihood of finding or not finding legally actionable public nuisance? Content analysis is well suited to answering this question in a body of case law that otherwise might appear unfathomable.

The Circularity of Facts in Judicial Opinions. Naturally, correlation does not equate with causation. This is especially so in analyzing the content of judicial reasoning, considering the serious circularity problem that judges marshal the facts and reasons that support the outcome of the case. Therefore, their opinions might not fully or accurately describe the real world facts or the true nature of the judge's decision process. Indeed, there is every reason to think just the opposite.

This limitation entails two distinct problems: factual incompleteness, and factual distortion. Incompleteness results because judges' presentations are meant only to explain as much of the facts as are necessary to justify the outcome. This judicial parsimony can severely distort analysts' measurement of facts that might be important across a range of cases. An apt example is the study of racial factors. Whites might be identified in only a fraction of cases where race is mentioned, but most likely this is because courts usually do not consider it appropriate to mention race unless they think this might be legally relevant in a particular case.

The second problem is the possibility that judges distort the facts they report to justify the results they reach. This is a highly contentious charge, but distortion does not have to amount to outright misrepresentation. Instead, distortion arises simply from the inevitability that courts select and filter the facts as relevant to the explanation of their decision, but doing only that creates a serious methodologic challenge, since it is circular to predict judicial outcomes from facts that reflect rather than generate the result.

Answering the Skeptics. There are four possible responses to judicial skeptics. First, scientific data aims only to be a reasonable approximation of underlying reality. As a probabilistic endeavor, it can tolerate a degree of imprecision, especially when such imprecision is randomly distributed, not reflecting biased measurement (for example, where the measurement of a particular dimension of law systematically underestimates or overestimates). Similarly, for facts reported by judges, even though they may not be a full account of the "real facts," they may be as close an approximation as is reasonably available to study a particular question. This assumption is not heroic. The lawyers and law professors who stake their life's work on believing (by and large) judges' renditions of facts are, on the whole, hardly naive idealists.

Second, researchers can specifically examine the fidelity of reported facts, looking for indications of distortion or incompleteness, to determine if the facts are close enough to reality for use in statistical analysis. One such technique is to compare facts reported in an appellate opinion with those reported in either the trial court's opinion or a dissenting opinion.

Third, the "bias" created by courts' justifying their decisions may be precisely what a researcher wishes to study. After all, the facts and reasons the judge selects are the substance of the opinion that creates law and binding precedent, so they merit careful study for this very reason. This justification calls, however, for precision in setting the goals of study. Instead of predicting outcomes, content analysis can aim simply at studying judicial reasoning itself, retrospectively.

Finally, the fact that content analysis may not provide definite answers to factors affecting judicial decisions does not mean the method lacks all value. Even if doubts remain about cause-and-effect relationships with judicial decisions, identifying apparent or possible associations of interest can merit further study using additional, and perhaps more experimental, methods.

Exploring the Landscape of Case law

Rather than trying to predict or explain case outcomes, content analysts can take advantage of the factual, rhetorical, and legal details in judicial opinions simply to describe or explore a body of case law. Observing and documenting what can be found in case law is more akin to mapping than to testing.   Like a naturalist exploring new (or familiar) terrain, researchers can code cases to document trends in the case law and the factors that appear important to case outcomes, such as the apparent effect of a new precedent, statute, or legal doctrine.

Wright and Huck (2002), for instance, code and analyze 440 decisions regarding milk production and purity standards during the 80 year period starting in 1860, in order to explore the historical question of whether courts were hostile or receptive to state legislatures' progressive public health agendas. They conclude that judicial hostility was greater than legal and social historians frequently recognize.

The primary criticism of some descriptive or exploratory studies is that they can draw conclusions about features of the legal landscape that cannot be observed fully from judicial opinions. As discussed more below, win/loss records from published opinions do not necessarily tell us about legal disputes that were never filed in court, or those that the parties settled, or those that judges resolved without written or published opinions. Nevertheless, even if judicial opinions offer a skewed view of what occurs elsewhere in the legal system, they are a highly valuable source for systematic study because they reveal the portion of the legal world that in many ways is most important. It is published opinions that set legal precedent and that guide lawyers.

Published opinions are especially probative of questions about the spread of ideas within the legal system or the types of information that judges appear to rely on. For instance, a number of studies analyze courts' reliance on different types of social science evidence (Hall & Wright, 2008). Naturally, some caution is warranted in concluding that a mention of some source in an opinion indicates the actual importance judges place on this type of evidence and argument. Still, with appropriate caveats on the claims being made, systematic study of how judges reason in their written decisions is perhaps the most compelling application of case content analysis because it best fits the method with the type of question that researchers are asking.

Finally, because published opinions represent "law," the amount, nature, and legal influence of particular dimensions of such law may well affect a diverse set of public health outcomes. To empirically study the public health effects of law, we need counts, weights, scales and other numeric indices of such law. Precise and specified procedures, and consistent implementation of such protocols, are required to meet scientific standards for reliable measurement.

Guidelines for Identifying and Coding Case Law

Assuming the decision has been made to conduct content analyses of case law (in contrast to traditional legal analysis), we next consider how best to design and implement such a content analysis. In brief, a content analyst selects a set of opinions on a particular subject via a pre-defined set of search and inclusion criteria, reads the documents systematically, records features of each one in a consistent and reliable manner, and then draws inferences about the use and meaning of those documents.

Selecting Cases

The first decision in any case-coding project is which cases to select. There are two components to consider: sampling frame, and selection method. The sampling frame is the theoretical universe of all relevant cases, and the selection method determines which cases will actually be sampled and studied. For both dimensions, researchers should specify exactly the protocol used (databases, search terms, repeated review and correction cycles, etc.) so that it is fully understood and reproducible by others.

Sampling Frame and Biases. Frequently in empirical studies, it is not feasible to observe all or most members of a relevant population. Therefore, the potential biases introduced by sampling method ordinarily are a topic of considerable methodological attention, so that a study sample accurately represents the true population of interest. Fortunately, most studies of legal decisions can avoid this concern because the sampling frame contains a small enough number of cases that universal sampling of all relevant cases is often feasible.

When the total population is too large to be manageable, however, sampling techniques might include: 1) true random sampling (best done by computer-generated list of random numbers); 2) systematic sampling, such as every fifth case; 3) quota sampling, such as all cases up to 200, per jurisdiction per year; or 4) purposive sampling, such as cases that are cited by leading treatises and casebooks or by other cases.

The more troubling question is the relevant sampling frame. What are the boundaries of the subject matter in question? Obviously, this depends critically on the study's central questions and purposes. Study questions can be narrowed to fit the sample frame that is available, or a theoretical sample frame can be imagined that is unrealistically broad, but that fits a more interesting or important set of questions the analyst wishes to pursue. Political scientists, for instance, often study political and institutional influences on judicial decisionmaking by looking, not at all Supreme Court cases or a random selection, but instead focusing on a particularly controversial or value-laden set of decisions, such as those involving freedom of speech or unreasonable searches and seizures.

Whenever the actual cases selected do not fully match the sampling frame that theoretically applies to the questions posed or studied, an issue of sampling bias exists. For instance, studies that sample cases until a certain date cannot, necessarily, claim with confidence that their findings reflect what happened after that date. Likewise for studies that sample from certain jurisdictions. Researchers should at least reflect on these potential distortions or limitations, and mention in their reports those that merit explanation.

One scholar, for instance (who happens to be the author of this monograph), chose to explore how courts determine the effectiveness of medical treatment in health insurance disputes, by studying all published judicial opinions resolving such disputes (Hall, Rust Smith, Naughton, & Ebbers, 1996). That universe of observations might be relevant for a narrower question, such as how appellate courts reason their decisions such issues, but the sample frame of all published opinions does not fully reflect what all courts do or how state trial courts actually make their decisions.

There is potential selection bias at each of many points in the litigation process. Only some human interactions produce disputes, only some disputes result in legal claims, many claims are settled, and many trial decisions are not appealed. Appellate courts regularly dispose of cases without opinions or decide not to publish some opinions, and computer databases inconsistently include cases that are not officially published. At each of these junctures, there are a variety of factors that potentially distort what one stage can reveal about the other. These biases can fundamentally threaten the validity or generalizeability of a study's findings. In these situations, careful consideration of selection biases may lead to major redesign of a study as originally conceived.

Sometimes, however, agonized hand-wringing can be minimized or avoided. No concern arises if the researcher defines the research question in terms that match the population of cases actually sampled. For instance, if it is precedential law that one wants to study, rather than simply the generalized behavior or attitudes of judges, then unpublished opinions are irrelevant and so excluding them requires no justification. In other situations, where excluded cases are theoretically relevant, the exclusion can easily be justified if the likely direction of bias or distortion is considered. When the bias runs in the same general direction as the study's findings (that is, the excluded cases are even more likely to exhibit the observed pattern), then including the additional cases would likely only strengthen the findings. The only major harm from excluding them is potentially to have missed some additional findings of interest or to have produced a false observation of no effect.

In other situations, likely differences between studied cases and omitted cases are sufficiently inconsequential that the omission should create no more a concern than other limitations obvious and inherent in the sample frame itself, such as one date range rather than another. All empirical studies are imperfect—observational (non-experimental) studies especially. The realistic standard for selecting cases is not a perfect match between sample frame and research objectives, but only a strong connection between the two.

Selection Techniques and Replicability. An essential attribute of scientific objectivity is the ability, at least in theory, to reproduce a project's findings using the same methods. Replicability is the overriding reason for using systematic content analysis, and transparency via reporting exact protocols used is a prerequisite for replicability. This is what confers scientific status on the findings and conclusions.

One component of replication is the method of selecting cases for study. Usually, this consists simply of specifying the particular structured search terms used to locate candidate cases in the Westlaw or LEXIS databases. However, mechanized searches are rarely refined enough to narrow the sample to only or mostly relevant cases. Cases that mention a topic of interest often do so only in passing. Those that decide an issue sometimes do so on technical or procedural grounds that are not relevant to a particular study. Therefore, further narrowing is usually needed in order to reduce an initial selection of candidate cases to those that are directly relevant to the research question.

Most legal researchers do so using somewhat subjective criteria of relevance that cannot be fully replicated. Another option, however, is to refine the initial mechanical search strategy. Useful strategies include searching case digests or headnotes rather than the full case itself, or searching a sample of cases selected initially because they cited particular statutes, or because they appear in a subject matter classification drawn by someone else, such as West's key numbering system or the publisher of a subject-matter-specific reporter. In effect, such researchers are relying on case selection criteria employed by someone else to establish probable relevance of cases.

Verifying the replicability of case selection is essential for a rigorous study, eliminating the possibility that a researcher subconsciously chose cases according to whether they appear to support the researcher's preliminary hunches. Either formal reliability testing, or case selection by someone who is otherwise uninvolved in the study, are ways to guard against this potential bias.

Coding Cases

Once cases are selected, a defined coding scheme focuses attention systematically on various elements of cases, and is a check against looking, either consciously or not, mainly for confirmation of predetermined positions. This effort to articulate beforehand the features of a case worth studying also allows researchers to delegate some or all of the reading to assistants. More importantly, coding cases, even for just qualitative description and analysis, strengthens the objectivity and reproducibility of case law interpretation. Experts in content analysis outline four basic steps that should be followed in coding any material (Krippendorff, 2004; Neuendorf, 2007):

1. Based on the questions most germane to the study, create a tentative set of coding categories a priori. After thorough evaluation, including feedback from colleagues, study team members or expert consultants, refine these categories.
2. Write a coding sheet and set of coding instructions (called a "codebook"), and train coders to apply these to a sample of the material to be coded. Pilot test the reliability (consistency) among coders by having multiple people independently code some of the material, and calculate the correlation across coders (i.e. inter-rater reliability).
3. Add, delete, or revise coding categories based on this pilot experience, and repeat reliability testing and coder training as required.
4. When the codebook is finalized, apply it to all the material. Then, or during that process, conduct a final, formal reliability test.

This section elaborates on each of these steps.

Coding Categories and Instructions. Categories used to code content of judicial decisions are tremendously diverse, owing to the wide range of questions that researchers pursue. Commonly used factors might be sorted into four general groups: 1) the parties' identities and attributes, 2) the types of legal issues raised and in what circumstances, 3) the basic outcome of the case or issue, and 4) the bases for decision. Coders often do not distinguish the "facts" of a case from various arguments that are made. Instead, they usually code simply for whether a variety of factual or legal factors are present in the case in some fashion. Coders should consider whether it suffices if these factors are merely alleged, realizing that the allegations may be sharply contested. If mere allegations are not sufficient, what is? Obviously, the procedural posture of a case (summary judgment versus post-trial) can complicate this evaluation.

Regarding the bases for decision, coders frequently distinguish between procedural and substantive rulings and they record the various types of authorities that courts cite or rely upon. Some researchers also code for the degree of importance that various factual or legal factors have in the court's analysis or holding. A common focus of coding is also the court's style of analysis or approach to statutory or constitutional interpretation, categorized in various ways.

Coding is not restricted to manifest variables that are explicit in the text; it has been shown to work well also for some "latent" variables that require inference or evaluative judgment. For instance, Johnson (1987) demonstrate the ability of law students to code cases with some degree of reliability for the clarity, complexity, and completeness of their discussion of facts, issues, holding, reasoning, and the law.

Coding experts advise researchers to create more coding categories, and to make coding more fine-grained, than the categories they may ultimately use. Even though this produces more information than the project will eventually require, the advantage is allowing the researcher to test different categorization schemes to learn through trial and error which work best. Ultimately, the goal is to maximize the exhaustiveness of coding, while keeping mutually exclusive categories—in other words, to capture all the relevant information, but to avoid having categories that duplicate or overlap each other. This does not mean, however, that a coding category must be devised for each possible nuance of relevance. Instead, categories should be used only if they occur with some frequency, or if the objective is to document their absence. Otherwise, rare or unusual features can be coded simply with a miscellaneous "other" option.

A good example of exhaustive and mutually exclusive coding is categorizing case outcomes. It is usually not a simple matter to define what counts as a win or loss across a range of cases. Appellate cases arise in a variety of procedural postures, they usually involve multiple issues, and each issue can be resolved in several different ways. Case coding projects often have to devise complex categories to capture all the relevant detail. The United States Court of Appeals Database, for instance, defines all possible case outcomes using nine categories (Songer, 2007). This illustrates that it is a better practice to be over inclusive at the coding stage, waiting until the analysis stage to collapse the various categories into discrete win/loss columns.

When categories are finalized, it is essential to good coding practice to record their description and specific instructions for their application. Obviously, this is necessary if coding is done by someone other than the researcher, such as student assistants. Even if authors do their own coding, the scientific standard of replicability requires a clear written record of how categories were defined and applied.

Experienced coders advise that errors will be reduced if coding forms are designed to minimize writing. For instance, a form might provide a checklist of factors to indicate presence or absence by ticking boxes rather than having to write in a number or letter. Also, while the objective is to reduce the need for coder judgment, detailed instructions can be conveyed either through the coding form itself, or in a supplemental manual. A balance should be struck between a form that is so spare it offers almost no on-the-spot instructional information for coders, forcing them to refer frequently to the detailed coding manual, versus a coding form that is overlong because each form contains a full set of instructions. Thus, it typically does not help to extensively revise succinct, well-written coding categories simply to satisfy the whim of each coder who might ask for more detailed instructions. It is inevitable that some measure of ambiguity will remain in how coding categories apply to atypical cases.

Choosing and Training Coders. A major dilemma in coding cases is whether principal investigators should do this work themselves, or instead whether they should supervise students (or others). In theory, the most scientifically rigorous method is for researchers to train others to do the coding and for coders to work completely independently once they are trained. Using generic coders ensures that the researchers' preliminary hypotheses and personal views do not bias the coding too much. Also, this can save researchers considerable time and effort in large coding projects. Moreover, the imposed discipline of training and supervising coders ensures that coding instructions are written in a way that others can follow. Training and using multiple coders promotes the reproducibility that is essential for good science.

Coding by law students is appropriate when some general legal knowledge is required but it is not necessary to be an expert in the field of study. Still, coding reliability improves the more that coders are trained. Researchers should describe how training was done in sufficient detail that others can replicate all of the steps.

Other considerations might counsel doing one's own coding, however. Training coders to achieve accurate and reliable results can be a difficult and time-consuming undertaking, one that may require considerably more resources and effort than researchers simply doing their own coding, especially in smaller projects. A relevant selection of cases is often sufficiently small that a single reader can handle the coding alone. Also, even trained coders can make a surprising number of mistakes, even on seemingly simple and objective criteria such as dates. Although delegating coding may promote reliability, this can threaten the validity of results if the information that coders record is not accurate or is too "dumbed down" to be meaningful.   It may be that student coders lack the level of expertise needed to code reliably the more complex or subtle, yet more meaningful, aspects of judicial opinions. If so, researchers will be sorely tempted to do their own coding. When this is done, however, it is especially important to conduct reliability tests by recruiting a colleague with similar expertise to independently double code at least a subset of cases.

Testing Reliability. Demonstrating the reliability of coding is an essential aspect of good content analysis. If coding categories are so objective and straightforward that it is obvious they can be applied consistently, then perhaps this step is not necessary, though that is rarely the case. If there are significant elements of subjectivity or uncertainty in applying coding categories to legal decisions, scientific rigor requires evaluation of whether different people would code the documents consistently. This is essential because the theory of coding—the reason systematic content analysis is done at all—is the implicit claim of reproducibility, that other researchers using the same methods will achieve approximately the same results. This claim is undermined if coding reflects primarily the subjective, idiosyncratic interpretation of the particular individuals who read the cases, or if coding has large elements of error or arbitrariness.

It is true that even without any reliability testing it is perfectly possible that a coding scheme in fact is reliable. But this cannot be assured unless investigators test coding reliability in some fashion. The best method is to conduct formal reliability tests during (at least) two stages in the process: initially, while piloting the draft coding process, and later, once coding categories and instructions are optimized.   Formal testing calls for at least two coders independently to code a sample of cases and to compare their results statistically.

The most common statistic is simple percent of agreement. However, a simple percentage does not account for the level of agreement that would be expected purely by chance. Because chance agreement varies according to the type of coding scheme (that is, a variable with two possible answers will naturally produce more agreement than a variable with eight possible answers), the best practice is to report one of several coefficients that reflect the extent of agreement beyond what is expected by chance. There are several such statistical tests, the most common of which is known as "Cohen's kappa" (after its inventor) or simply the kappa statistic. Ranging from 0 to 1, kappa indicates the proportion of observed agreement that exceeds what would be expected by chance alone, with 0 indicating agreement entirely by chance and 1 indicating perfect agreement.

If statistics such as these are used, they must be employed correctly. One mistake is to test the overall reliability of all variables combined. The correct method is to test and report each variable's reliability because reliability can vary widely across items and aggregate statistics can mask serious problems with key variables. Also, when the response pattern for a variable is highly skewed (say, one of several available responses occurs much more frequently than the others), this should be noted or taken into account in the statistical measure used. Otherwise, the nominal level of agreement can be deceptive. If one were to code for the presence or absence of 100 factors in each case, most likely only a dozen or so will appear in any one case. Testing for coding reliability may find a very high percent of agreement, then, but only because most factors are not present in most cases. The key question, though, is whether coders agree when they indicate a factor is present.

When reliability testing reveals discrepancies, as it almost always will, this will usually point to unresolved questions in the coding instructions, problems that can be corrected if the error appears after the pilot phase rather than after the completed coding. Also, poor reliability in the pilot round of coding often reveals conceptual ambiguity that can be clarified to more accurately measure the dimensions or their components that are actually most relevant to the particular research question.

After final coding, compulsive researchers might try to get to the bottom of remaining disagreements and resolve all discrepancies, both in the reliability testing sample and across the entire selection. When there are large numbers of cases being coded, resolving every discrepancy may be unnecessary and impractical. Disagreements sometimes arise from overt errors, but often they result simply from judgment calls or inevitable ambiguities that may be virtually impossible to eliminate without compromising the independence of individual coders. Perfect reliability is the goal, but rarely fully achieved. A key requirement of science is transparency—reporting the exact levels of reliability of the resulting data.

Refining coding rules to eliminate all elements of ambiguity is usually not possible, no matter how prescriptive the rules. Plus, each time the rules are rewritten, the best practice would be to retest the refined rules for reliability, producing a never-ending cycle in search of elusive perfection. Therefore, coders should learn to live with a certain degree of imperfection once coding is found to be reasonably reliable, and draw appropriately modest conclusions when relying on variables with weaker levels of inter-coder reliability.

Although there is broad agreement on the desirability of testing for reliability, and some agreement on the methods for doing so, there is not firm agreement on what level of reliability is acceptable. The goal is aspirational — to achieve high levels of agreement — rather than merely to rise somewhat above purely random agreement. One suggested rule of thumb is that a reliability coefficient of .8 (that is, data agree 80% more than mere chance agreement) is good, with indices from .67-.8 being sufficient for "tentative conclusions" (Krippendorff, 2004). Others claim that this is too demanding, especially for coding categories that produce more skewed responses, since even small levels of disagreement can cause the statistical index to drop rapidly. Therefore, other methodologists provide a more lenient classification for the kappa statistic (Lombard, Snyder-Duch, & Campanella Bracken, 2005): <0.00 is Poor; 0.00-0.20 is Slight agreement; 0.21-0.40 is Fair; 0.41-0.60 is Moderate; 0.61-0.80 is Substantial; and 0.81-1.00 is Almost Perfect. Keep in mind that these recommendations are for agreement levels beyond what is expected by chance. For raw, unadjusted percents, agreement levels below 70-80% are usually not considered to be good.

If coder agreement is not acceptable, researchers must either re-train coders, revise their coding categories, decide not to use the data, or use the data but with appropriate caveats. Following best practices, the first two options call for re-testing of reliability. One convenient remedy is to combine marginally-reliable detailed coding into a more aggregated category that has better reliability.

Alternative Coding Techniques. Researchers might consider alternatives to independent coding by assistants. One is to have a group of assistants code each case and then assign the value that is coded by the majority. Group coding creates the impression of greater objectivity, and may in fact improve reliability, but this is not necessarily the case. Resolving split votes with whatever the third person thinks might be as arbitrary as using a single coder. The only way to find out for sure is to test the reliability of panel coding by coding a sample of cases independently with a different panel.

Similarly, some researchers have coders confer when they disagree in order to seek consensus, or the researcher uses her own expertise to resolved disagreements. Again, this may or may not improve reliability, but it does not establish reliability. The process of reaching consensus might be arbitrary, or the lead investigator's expert view might not be objectively reproducible.

A variation of these techniques is expert panel consensus. Developed for evaluating medical judgments, this has not been used so far for legal judgments but it is worth exploring. Following what is known as the Delphi technique, each expert first rates a case independently, then learns how peer experts have rated it, and then, following discussion, each expert gives an independent final rating, with the majority controlling when there is not unanimity (Shekelle et al., 1998). This has been shown to be a fairly reliable method for rating highly complex and judgmental aspects of medical decisionmaking (Park et al., 1986). It combines elements of "gold standard" expertise with consensus building and majority rule.

Finally, there is an innovative technique that avoids altogether the vagaries of training coders and demonstrating reliability: using completely mechanical forms of content analysis that can be done by computer or simple computation. For instance, some studies count the number of words or paragraphs devoted to discussing particular factors as an indication of the factors' relative importance. Also interesting is research that analyzes judicial texts entirely by computer, looking for revealing patterns in syntax or semantics (McGuire, Vanberg, Smith, & Caldeira, 2009; Wahlbeck, Spriggs, & Sigelman, 2002).

Analyzing Cases

A credible content analyst does not necessarily need to use complex or sophisticated statistics—or, indeed, any statistics at all. Often, researchers simply report counts and frequencies to show how commonly a given feature appears in the cases. Quantitative descriptive analyses may be sufficient to document trends in the case law, to challenge conventional wisdom, or to raise provocative questions meriting further study. Because many case-counting studies code the entire universe of relevant cases, statistics are not essential for analyzing the probability that the sample cases reflect the reality in a larger population.

Moreover, content analysis need not involve numbers at all. Instead, it can employ rigorous methods of purely qualitative analysis that focus on themes and patterns that are best understood through conceptual description and narrative illustrations rather than numbers. To empirically evaluate the public health effects of case law, however, requires the use of counts and numeric indices and scales, and so statistical analyses are often essential.

One danger in using statistical testing in exploratory studies is that, without a tightly controlled analytical focus, such as a predefined set of hypotheses that are being tested, it becomes too easy to find associations and patterns of apparent significance entirely by chance. If enough variables are examined and enough comparisons are made, odds are that significant findings will emerge, but some or all of these apparent findings could be due entirely to chance without additional statistical adjustments for the number of possibilities that were explored.

Potentially more revealing is multiple regression analysis, which can uncover hidden relationships among multiple factors, both internally within decisions or between decisions and external factors. In attempting to explain the legal outcomes or health-related effects of a set of cases, for example, several factors may each appear significant by themselves, but when each is held constant, only one or two factors may emerge as the most important predictors of decisions. Sometimes, factors that legal analysts thought were dominant or important turn out to be red herrings. Alternatively, factors that, standing alone, may not appear significant might emerge as such once the influence of other factors is controlled statistically. It is advisable, however, to use regression analysis only for cases that are relatively homogenous, focusing on a single or narrow set of legal issues. Otherwise it may become too difficult to measure and control for all the relevant variables.

Another aspect of statistical analysis worth considering in broad perspective is whether each case (or part of a case) should be given equal weight. It is possible to weight each case according to an objective measure of its significance, such as how often it has been cited or followed, or where it stands in a line of precedent. The difficulty with this approach, however, is deciding how much weight to assign. Nascent efforts to apply network analysis to the citation patterns among cases may eventually prove fruitful in assigning appropriate weights to different cases (Smith, 2005). But, absent any objective means to assign different quantitative weights, the best option is to classify cases qualitatively into different categories and analyze each separately—such as major versus minor decisions, or leading versus following decisions.

A final concern is the appropriate unit of analysis. Rather than each case counting the same, case law can be grouped into separate jurisdictions in order to assign a legal rule to each location. Doing that enables exploration of how the adoption of particular rules of law relate to other occurrences include health outcomes.

Conclusion

Content analysis is a valuable research tool for documenting what courts do and what they say. The insights gained from uniform content analysis of large numbers of opinions supplement the deeper understanding of individual opinions that comes from traditional interpretive techniques. The content of judicial opinions can be important in the study of the broader social, economic, and political systems that interact with judicial precedent, but cases are also well worth scientific study in their own right.

The major limitation of content analysis (a limit that applies equally to traditional interpretive methods) is that facts and reasons given in opinions cannot be treated as accurate and complete. Therefore, researchers should be cautious about the meanings they attach to observations made through content analysis. Within these bounds, content analysis is well suited to studying the connections between judicial opinions and other parts of the social, political, or economic landscape.

Scientific methods complement conventional legal research methods in three key ways. Content analysis can verify or refute descriptions of case law that are based on more anecdotal or subjective study. Second, content analysis can identify surface patterns (which are sometimes hidden from the naked eye), to be explored more deeply through interpretive, theoretical, or normative legal analysis. Third, systematic numeric content coding of case law opens up major new avenues of research to better understand the many ways in which law affects population health.

References

Friedman, B. (2006). Taking law seriously. *Perspectives on Politics, 4*(2), 261-276.

Goldsmith, J., & Vermeule, A. (2002). Empirical methodology and legal scholarship. *The University of Chicago Law Review, 69*(1), 153-167.

Haar, C. M., Sawyer, J. P., Jr., & Cummings, S. J. (1977). Computer power and legal reasoning: A case study of judicial decision prediction in zoning amendment cases. *American Bar Foundation Research Journal, 2*(3), 651-768.

Hall, M. A., Rust Smith, T., Naughton, M., & Ebbers, A. (1996). Judicial protection of managed care consumers: An empirical study of insurance coverage disputes. *Seton Hall Law Review, 26*, 1055-1068.

Hall, M. A., & Wright, R. F. (2008). Systematic content analysis of judicial opinions. *California Law Review, 96*(1), 63-122.

Hoffman, D. A., Izenman, A. J., & Lidicker, J. R. (2007). Docketology, district courts, and doctrine. *Washington University Law Review, 85*(4), 681-752.

Johnson, C. A. (1987). Law, politics, and judicial decision making: Lower federal court uses of supreme court decisions. *Law & Society Review, 21*(2), 325-340.

Krippendorff, K. (2004). *Content analysis: An introduction to its methodology* (2nd ed.). Thousand Oaks, CA: Sage.

Loevinger, L. (1961). Jurimetrics: Science and prediction in the field of law. *Minnesota Law Review, 46*, 255-275.

Lombard, M., Snyder-Duch, J., & Campanella Bracken, C. (2005). Practical resources for assessing and reporting intercoder reliability in content analysis research projects. Retrieved July 1, 2011, from http://www.temple.edu/mmc/reliability/

McChesney, F. S. (1993). Doctrinal analysis and statistical modeling in law: The case of defective incorporation. *Washington University Law Quarterly, 71*(3), 493-534.

McGuire, K. T., Vanberg, G., Smith, C. E., & Caldeira, G. A. (2009). Measuring policy content on the u.S. Supreme court. *The Journal of Politics, 71*(04), 1305-1321.

Neuendorf, K. A. (2007). *The content analysis guidebook*. Thousand Oaks, CA: Sage.

Park, R. E., Fink, A., Brook, R. H., Chassin, M. R., Kahn, K. L., Merrick, N. J., et al. (1986). Physician ratings of appropriate indications for six medical and surgical procedures. *American Journal of Public Health, 76*(7), 766-772.

Revesz, R. L. (1997). Environmental regulation, ideology, and the D.C. Circuit. *Virginia Law Review, 83*(8), 1717-1772.

Revesz, R. L. (2001). Congressional influence on judicial behavior? An empirical examination of challenges to agency action in the D.C. Circuit. *New York University Law Review, 76*, 1100-1137.

Shekelle, P. G., Kahan, J. P., Bernstein, S. J., Leape, L. L., Kamberg, C. J., & Park, R. E. (1998). The reproducibility of a method to identify the overuse and underuse of medical procedures. *New England Journal of Medicine, 338*(26), 1888-1895.

Smith, T. A. (2005). The web of law. SSRN eLibrary Retrieved 06-11, from http://ssrn.com/paper=642863

Songer, D. S. (2007). Extending the multi-user database of decisions of the U.S. Courts of appeals 1997-2002 [Dataset]. Retrieved October 1, 2011, from http://www.wmich.edu/nsf-coa/

Wahlbeck, P. J., Spriggs, J. F., & Sigelman, L. (2002). Ghostwriters on the court? *American Politics Research, 30*(2), 166-192.

Wright, R. F., & Huck, P. (2002). Counting cases about milk, our "most nearly perfect" food, 1860-1940. *Law & Society Review, 36*(1), 51-112.

Please cite this document as:

Hall, Mark A. (2011). Coding case law for public health law evaluation. *PHLR Methods Monograph Series.*