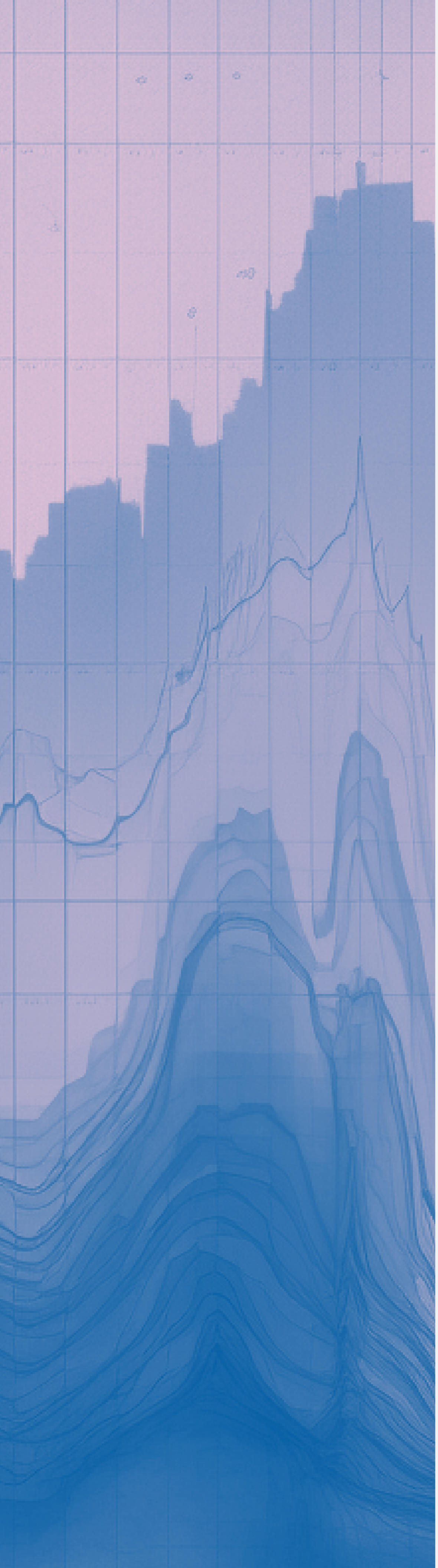
October 2023



THE FUTURE SOCIETY

IEEE

Model Protocol for Electronically Stored Information (ESI)

Guidelines for Practitioners

Note on the Preparation of these Documents

These documents grew out of conversations between Nicolas Economou and Bruce Hedin following the publication, in 2019, of the first edition of *Ethically Aligned Design*, the flagship publication of the IEEE's Global Initiative on Ethics of Autonomous and Intelligent Systems. Their conversations centered on the question of how to enable practitioners to act on the principles articulated in that document (to which both Mr. Economou and Dr. Hedin had contributed). Recognizing the benefits of approaching the question through a specific application, they focused on AI-enabled systems applied to legal discovery and sought to identify a mechanism that would enable the use of those systems to be grounded in an informed trust (and, more specifically, in accordance with IEEE's recommendations, grounded in the key principles of effectiveness, competence, transparency, and accountability). That discussion led to the vision of creating an ESI Protocol that would instantiate the IEEE trust principles.

Mr. Economou and Dr. Hedin initially shared their vision with interested parties from the IEEE (Konstantinos Karachalios), The Future Society (Nicolas Miailhe), and NYU's Center on Civil Justice (David Siffert, Arthur Miller, the late Peter Zimroth). Having received positive responses from those parties, they sought to gain additional validation of the relevance of their vision by outlining their ideas to a wider circle of stakeholders (a group of judges, practitioners, consultants, and academics with an interest in legal discovery, most of whom eventually became a standing review group for the documents).

After those initial steps, the project remained largely at the outline stage for a couple of years as other matters intervened. In 2022, however, thanks to support from the IEEE and further guidance from The Future Society and NYU's Center on Civil Justice, the project was resumed in earnest. Dr. Hedin took the lead in drafting; The Future Society's Samuel Curtis took on the role of project manager. In keeping with the initial vision, Dr. Hedin drafted three documents: a model protocol, a line-by-line commentary on that protocol, and a handbook for practitioners seeking a deeper understanding of the procedures prescribed in the protocol. The review group was revived in 2023 and generously gave their time to review multiple drafts of the documents. As the project approached the final draft stage, Hon. John M. Facciola, U.S. Magistrate Judge (Ret.) agreed to add a preface to the documents.

This is the story of the documents before you. As will be evident, there are many individuals to whom thanks are due; please see the *Acknowledgements* section for an effort at recognizing them. While the documents have undergone multiple reviews, it is to be expected that more will be learned when they are put into practice. Feedback, comments, and suggestions are welcome; please send them to the following addresses.

Bruce Hedin
Hedin B Consulting
bhedin@hedinb.com

Samuel Curtis
The Future Society
samuel.curtis@thefuturesociety.org

© 2023 by The Future Society and IEEE.
This work is licensed under a Creative Commons Attribution-Non Commercial 4.0 International License.

Table Of Contents

THE
FUTURE
SOCIETY

IEEE

Model Protocol for Electronically Stored Information (ESI)

Guidelines for Practitioners

September 2023

Introduction

The *Model ESI Protocol* ("Protocol") specifies procedures for validating the results of a review effort; the *Commentary* to that protocol ("Commentary") provides guidance on implementing those procedures. The Protocol and Commentary will substantially meet the needs of most practitioners. There are circumstances, however, such as when answering a question raised by opposing counsel or when adapting to a mid-review change in the data landscape, in which a deeper grounding in the concepts and methods that underly the approach to validation specified in the Protocol and a more detailed understanding of the steps to be followed in executing its procedures are required. The purpose of the *Guidelines for Practitioners* ("Guidelines") is to provide the additional level of depth required in such circumstances.

The intended audience of the Guidelines consists of two categories of stakeholders in legal discovery processes.

- **Advanced practitioners** seeking to execute the Protocol's provisions in a sound and confident manner. In this category are lawyers, consultants, and vendors who need the know-how to design, execute, and defend a sound validation exercise.

- **Stakeholders** seeking to understand the resources, time, and cost required to gain a well-grounded trust in the results of a review effort. In this category are parties, lawyers, and judges who, while not themselves engaged in the execution of validation procedures, do need, in order to arrive at well-informed answers to the questions they face, a basic understanding of what is required to obtain sound evidence of the effectiveness of a review.

To meet these needs, the document provides guidance on three topics: (1) the calculations required in estimation procedures, (2) the reasoning that is the basis for sample size selection, and (3) statistical terminology. A working familiarity with each of these is essential if a practitioner is to implement a validation protocol in a sound manner. Each topic is treated in a separate chapter; the specific organization is as follows.

- **Chapter 1: A Guide to the Estimation of Validation Metrics.** This chapter provides guidance on how to obtain sound estimates of validation metrics and on how to interpret the results. The chapter is organized into five sections.
    - **Section 1.1: Preliminaries.** On the conceptual foundations for the procedures discussed in subsequent sections.
    - **Section 1.2: On the Validation of Exclusionary Steps.** An overview of the calculations required to obtain measurements for validating steps (such as using search terms) taken to narrow the scope of data subject to review.
    - **Section 1.3: On the Validation of a Review Process.** An overview of the calculations required to obtain measurements for validating a review process.
    - **Section 1.4: Additional Circumstances and Metrics.** An overview of the calculations required to handle more complex cases (such as those involving phased productions) and the calculations required to obtain measurements supplementary to recall (such as precision and prevalence).
    - **Section 1.5: Worked Examples.** An introduction to the worked examples provided in Appendix B.
- **Chapter 2: A Guide to Sample Size Selection.** This chapter provides guidance on the size of samples to use in the various validation exercises that will be occasioned in the course of a review. It is organized into two sections.
    - **Section 2.1: On the Size of Samples Other than the Recall Negative Sample.** Guidance on the size of samples to use in validating exclusionary steps (both Positive and Negative Samples) and on the size of the Positive Sample used in the estimation of recall.
    - **Section 2.2: On the Size of the Recall Negative Sample.** Guidance on the size of the Negative Sample used in the estimation of recall. This section is the primary focus of the chapter and covers (i) methodology for analyzing the power of a sample, (ii) setting a criterion and finding a candidate sample size, and (iii) further analysis of candidate sizes.
- **Chapter 3: A Glossary of Terms of Art Used in Validation.** This chapter provides guidance on terminology used in discussions of sampling and measurement. It is not intended as a comprehensive glossary of ESI or e-discovery; its focus is on terms of art used in discussions of validation.[1]
- **Appendix A: Equation Library.** This appendix is a reference list of the equations that are used to obtain estimates (and associated margins of error) of the metrics used in the validation of review processes. It is used primarily in conjunction with the discussion of procedures in Chapter 1.
- **Appendix B: Worked Examples.** In this appendix, we walk through four examples of applying the procedures discussed in Chapter 1. The purpose is to provide practitioners with the opportunity to strengthen their familiarity with the required calculations by walking through specific examples.

[1] The first instance of any term included in the glossary will be linked to its glossary entry via hyperlink (indicated by double underlining).

Chapter 1: A Guide to the Estimation of Validation Metrics

Section 1.1: Preliminaries

This chapter describes procedures for executing the calculations required to arrive at the statistical estimates (and their associated margins of error) used in the validation of review processes. Making meaningful use of statistical estimates, however, is about more than simply executing a series of calculations. Making effective use of the results of statistical procedures requires an understanding of why we set out to execute those procedures to begin with as well as an understanding of the limitations of statistical estimation. This section discusses these preliminary considerations, covering, more specifically, the following topics: the goals of a validation exercise; the metrics best suited to meeting those goals; the need for a sample-based approach and the limitations inherent in that approach. The section concludes with a note on the organization of the discussion.[2]

Goals

The goal of a validation exercise is to generate data that will give empirical grounding to an assessment of whether a review process has met its intended objective. It is important to note, however, that the "intended objective" of the process under scrutiny is not defined by the validation exercise; the intended objective is a matter of law and is defined by the requirements, conditions, and norms that prompt the execution of the process. In the case of an ESI Review, these objectives are defined by rules of procedure and by case law and are colored by considerations of what is "reasonable," what counts as "good-faith" effort, and what is "proportionate" in a given set of circumstances. Assessing whether a review process has met its objective will therefore require not only the evaluation of the data supplied by the validation exercise but also the application of legal judgment.

Metrics

Allowing for the fact that the goal of a validation exercise is to *support*, not *make*, a legal judgment, a good validation exercise will be one that supplies the data that is most telling in an assessment of the effectiveness of the process under scrutiny. Those data will typically take the form of measurements.[3] In the case of the protocol that is the occasion for these guidelines, there are two types of processes that are subject to validation: exclusionary processes used to narrow the set of documents subject to a review for responsiveness and the review process itself. As discussed in greater detail below,[4] for the former type of process, the most telling metric is a comparison of the total number of responsive documents included in the downstream review to the total number excluded; for the latter process, the most telling metric is recall.[5]

Sampling

[2] For another helpful discussion of concepts and procedures involved in evaluating the effectiveness of review processes, see Lewis 2016.

[3] For a helpful discussion of metrics useful in evaluating review processes, see Webber and Oard 2016.

[4] See the subsection *Interpreting the numbers* in Section 1.2*: On the Validation of Exclusionary Steps* and the subsection with the same heading (*Interpreting the numbers*) in Section 1.3*: On the Validation of a Review Process*.

[5] This is not to say that other metrics do not add valuable context or color to the perspective provided by recall; they may and, in the detailed procedures given below, we also cover the calculation of two such metrics: precision and prevalence. It is also worth emphasizing, in any discussion of metrics, that numbers do not tell the whole story; hence the need for the metrics to be supplemented with qualitative analysis (see below).

The document populations that are the domain of modern discovery requests are, at least in the cases in which advanced review technologies are applied (and in which validation issues become especially acute), quite large, generally exceeding the limits of what could reasonably be assessed by exhaustive manual review. This means that, when we set about obtaining the metrics required for the validation of a process, we will almost always be seeking, not the true value of the metric (which could be obtained only via an exhaustive canvassing of every item in the population, and so is an impractical goal[6]), but a sample-based estimate of the true value. From this fact follow some considerations that must be kept in mind, both when designing a validation exercise and when assessing the results of one.

- There will always be some uncertainty associated with a statistical estimate. There is an element of chance in the selection of the sample that is used as the basis for the estimate, and the operation of chance may make a sample-based estimate higher or lower than the true value.
- It is important to gauge the amount of uncertainty associated with an estimate. The science of statistics allows us to quantify, through the calculation of a margin of error or confidence interval, the possible impact the operation of chance can have on our estimates.
- Increases in sample size can reduce the amount of uncertainty associated with an estimate (i.e., reduce the margin of error associated with the estimate).
- There are limits, however, on what can be accomplished by increases in sample size, so practitioners must always balance the trade-off between information gained and the cost of gaining it.[7] There is a point at which the diminishing returns of further increases in sample size will make it impractical to seek further reductions in the margin of error associated with an estimate. In the case of recall estimation, these challenges are particularly acute when the prevalence of responsive material in the review population is low.[8]

If these limitations are given due consideration, sampling[9] can be a powerful tool for obtaining the information required to validate the results of a review process.

Qualitative analysis

Finally, as provided for in the Protocol and underlined in these guidelines, it must be remembered that numbers do not tell the whole story: it is important to supplement quantitative measures (such as estimates of recall) with qualitative analysis (such as an assessment of the uniqueness and importance of any responsive documents found to have been missed by the review process).

[6] It is an impractical goal because the resources required would be prohibitive. Moreover, even if the resources were available, eliminating the human error from such a review (and knowing that you had eliminated all such error) would, for all practical purposes, be impossible.

[7] And doing so will often also introduce the legal concept of proportionality.

[8] When prevalence is extremely low, it may not be possible to find a sample size that meets both the requirement of being practically manageable and that of reducing the margin of error to a meaningful size. In such cases, alternative approaches to gathering empirical evidence of effectiveness must be considered. For more on the relation between prevalence and sample size, see Chapter 2.

[9] Throughout these guidelines, we assume that the method of selecting samples, whether from the full population or from subsets of the population, is, as specified in the Protocol, simple random sampling.

Note on the presentation of quantitative procedures

Sound validation of the processes involved in responding to a request for production requires measurement. Measurement, especially sample-based measurement, requires math: in order to obtain the required metrics, we must execute certain mathematical operations. The operations required for the metrics we seek are not, however, complex (generally not going beyond the elementary operations of addition, subtraction, multiplication, division, and exponentiation) and the specific formulae in which those operations figure are not very numerous (the same formulae are used multiple times in the derivation of results). For practitioners, even those without a deep background in mathematics or statistics, the steps required to obtain meaningful measures are always within reach; what is needed is care and attention to detail.[10]

In presenting the steps, the organization we adopt is as follows. In Appendix A (the "Equation Library") we list notational conventions used in the discission and all the equations used to obtain estimates (and associated margins of error) of the metrics used in validating both exclusionary steps and review processes. In the procedural sections that follow this one (Sections 1.2, 1.3, and 1.4), we specify the steps to be followed in obtaining the salient metrics, invoking, when needed, the equations listed in the appendix.[11] Section 1.5 introduces some worked examples that are provided in Appendix B; the intent of the examples is to show how the calculations discussed in this chapter are applied in practice and to allow practitioners to deepen their understanding of the procedures by trying the examples on their own.

Section 1.2: On the Validation of Exclusionary Steps

When an exclusionary step is taken for the purpose of reducing the amount of non-responsive data to be included in the Review Set (e.g., when search terms are developed and applied for the purpose of "culling" the collected data), the validation of the step's effectiveness boils down, in terms of the *quantitative* component of the validation exercise,[12] to a comparison between two numbers: an estimate of the number of responsive documents in the set designated for inclusion in the downstream review and an estimate of the number of responsive documents in the set designated for exclusion from downstream review. Obtaining these estimates (and their associated margins of error) is, statistically speaking, a straightforward exercise in obtaining an estimate of the total number of items of interest in a single population, an exercise the procedures for which can be found in any textbook on sampling.[13] In our case, we need to execute the procedures twice, once for the *Positive Set* (the set of documents designated for inclusion in the downstream review) and once for the *Negative Set* (the set of documents designated for exclusion from downstream review). The specifics are as follows.

Inputs

To calculate an estimate of the total number of items of interest in a single population (and the margin of error associated with that estimate), we need three input numbers: (1) the size of the population that is the

[10] And, of course, practitioners must be cognizant of the assumptions underlying the statistical methods being applied. This, more than in the execution of calculations, is where expertise is required and is where practitioners may wish to call upon the support of consultants or other individuals with the appropriate scientific or statistical skills.
[11] When an equation is invoked in a procedural section, it is referenced by the right-margin numbering in Appendix A and made accessible via hyperlink.
[12] For more on the *qualitative* component of a validation exercise, see the Appendices A and B of the Protocol along with the associated discussion in the Commentary.
[13] For example: Thompson 2002: 16*f*.

domain of the estimation exercise, (2) the size of the sample we have drawn from that population, and (3) the number of items of interest that we observe in the sample. When validating an exclusionary step, we, as just noted, execute the procedures twice, once for the Positive Set and once for the Negative Set. We therefore have six input numbers in all:

- N_+: The number of documents in the Positive Set;
- n_+: The number of documents in the Positive Sample;
- r_+: The number of responsive documents observed in the Positive Sample;
- $N_\circ$: The number of documents in the Negative Set;
- $n_\circ$: The number of documents in the Negative Sample; and
- $r_\circ$: The number of responsive documents observed in the Negative Sample.

With regard to N_+ and $N_\circ$, it may be observed that these numbers can be obtained once the Positive and Negative Sets have been defined (i.e., once the exclusionary steps[14] under evaluation have been applied to the Collected Set). For example, in the case of search-term culling, the Positive and Negative Sets can be defined once the search terms have been developed to the point at which they are deemed ready for validation and have been applied, in aggregate, to the Collected Set.[15] In this case, the Positive Set is defined as the set of documents hit by at least one search term, together with any associated family members of such documents.[16] The Negative Set is defined as the remainder of documents in the Collected Set (i.e., the set of documents neither hit by a search term nor associated, by family relation, with a document hit by a search term).

With regard to n_+ and $n_\circ$, it may be observed that default specifications for these values (the sample sizes) are given in the Protocol: 400 for n_+ and, for $n_\circ$, either 6,000 (if validating an application of search terms) or 1,200 (if validating a metadata-based exclusion). As provided for in the Protocol, practitioners may depart from the default specifications when circumstances warrant; when Practitioners do so, n_+ and $n_\circ$ will of course represent the sizes of the samples actually drawn.

With regard to r_+ and $r_\circ$, it may be observed that, of the six input numbers, it is only r_+ and $r_\circ$ that require additional work beyond the application of the exclusionary step under evaluation. The validation samples

[14] As noted in the Commentary to the Protocol, a responding party may choose to conduct a single aggregate test of the result of applying multiple exclusionary steps (e.g., the result of applying multiple metadata-based exclusions or the result of both a metadata-based exclusion and a search-term-based exclusion). In the case of such an aggregate test, the Positive Set would be defined as the set of documents to be retained in the Review Set (together with any associated family members of such documents), **after applying all the exclusionary steps that are being evaluated in the exercise**. The Negative Set would then be simply the remainder of documents in the Collected Set. While such aggregate testing will provide actionable information, it should be noted that, in the case of metadata-based exclusions, tests focused on a specific exclusion may be more efficient and informative.

[15] Minus any documents to be excluded from the Review Set on other grounds (e.g., because of a metadata-based exclusion or because the document has characteristics, such as text deficiency, that make it unfit for search term filtering).

[16] We define the Positive Set to include family members of documents hit by a search term because those family members will be included in the Review Set (families are kept intact when creating the Review Set). What we are testing when we validate exclusionary processes is the net effect of applying those processes (i.e., what is included in the Review Set *vs*. what is not included in the Review Set), so as long as a document makes it into the Review Set, regardless of *how* it makes it in, it counts as in the Review Set (i.e., in the terms of the validation exercise, belongs in the Positive Set).

must be drawn and (manually) reviewed; once that is done, we will have the counts of the responsive documents observed in each of the validation samples.

Procedures

Once the input numbers are in hand, obtaining estimates and margins of error for the target metrics is simply a matter of applying the appropriate equations from the *Equation Library*.[17] These are as follows.

1) Obtain **point estimates** of the target metrics.
   a) Find the point estimate for the number of responsive documents in the Positive Set (t_+).
      i) Using as inputs n_+ and r_+, apply Equation 1 to obtain the estimated proportion of responsive documents in the Positive Set (p_+).
      ii) Using as inputs p_+ (the output of the preceding step) and N_+, apply Equation 3 to obtain the point estimate for the number of responsive documents in the Positive Set (t_+)
   b) Find the point estimate for the number of responsive documents in the Negative Set ($t_\circ$): repeat the steps specified under 1(a), replacing the Positive-Set inputs with the Negative-Set inputs ($N_\circ, n_\circ, r_\circ$).
2) Obtain the **margins of error** associated with the point estimates.
   a) Obtain the margin of error associated with the t_+ estimate ($M(t_+)$).
      i) Using as inputs N_+, n_+, and p_+, apply Equation 2 to obtain the estimated variance of the proportion estimator ($var(p_+)$).
      ii) Using as inputs $var(p_+)$ (the output of the preceding step) andN_+, apply Equation 4 to obtain the variance of the total estimator ($var(t_+)$).
      iii) Using as input $var(t_+)$ (the output of the preceding step), apply Equation 5 to obtain the margin of error associated with the estimate of the total number of responsive documents in Positive Set ($M(t_+)$).
   b) Obtain the margin of error associated with the $t_\circ$ estimate ($M(t_\circ)$): repeat the steps specified under 2(a), replacing the Positive-Set values with the Negative-Set values.
3) **Summarize** the result.
   a) Responsive documents in the set designated for further review: $t_+ \pm M(t_+)$.
   b) Responsive documents in the set to be excluded from further review: $t_\circ \pm M(t_\circ)$.

Interpreting the numbers

Our fundamental question, when we seek to validate the effectiveness of an exclusionary step, is whether the step has been effective at identifying a subset of the Collected Set that is largely void of responsive material and so can safely be excluded from the Review Set. In the case of search terms, for example, the

[17] While applying the equations is simply a matter of executing a sequence of elementary mathematical operations, some practitioners may still find it helpful, or reassuring, to engage a consultant with the appropriate expertise to provide support in carrying out this part of the validation exercise.

question is whether the search terms have been effective at sweeping (almost) all the responsive documents that reside in the Collected Set into one smaller subset (which we have called the *Positive Set*), leaving a second subset (which we have called the *Negative Set*) largely void of responsive documents. The metrics we obtain via the sampling and estimation exercise just reviewed are intended to help us answer this question. In this section, we provide further guidance on the metrics and on how to interpret them.

Why sample from the Positive Set? Considering the design of the validation exercise, a practitioner might reasonably ask: *Why do we have to sample from the Positive Set? After all, our concern, at this stage in the review process, is whether we are missing responsive documents. Any missed responsive documents will be found in the Negative Set, so won't it suffice to focus our sampling efforts there (and not on the Positive Set)?*

The answer is that we cannot assess the significance of any results we obtain from the Negative Set without the context provided by results from the Positive Set. Our sampling of the Negative Set will provide us with an estimate of the number of responsive documents that will be excluded by taking the step in question. That number alone, however, does not tell us whether we should be worried or confident about taking the exclusionary step. To answer that question, we need both that number *and* the estimate of the number of responsive documents that will be included in the Review Set. Only by putting "misses" (*false negatives*) in relation to "hits" (*true positives*)[18] will we be able to arrive at a meaningful assessment of the empirical grounds for the exclusionary step.[19] Hence our need to obtain an estimate of the true positives and hence, in turn, our need to sample from the Positive Set.

Why a number, not a percentage? Another question that a practitioner, considering the design of the validation exercise, might ask is: *Why do we need to obtain an estimate of the number of responsive documents in the Negative Set? Wouldn't it be easier simply to obtain an estimate of the percentage of responsive documents in that set? Wouldn't that have the advantage of allowing us to set a consistent threshold (e.g., a responsive rate of 1% or less) for evaluating the risk posed by an exclusionary step?*

The answer is that percentages, in isolation, can be misleading. A low percentage of a large population may still amount to a large number of documents, while a high percentage of a small population may amount to a small number of documents. In legal discovery, when, for example, we are using search terms to reduce the size of the set subject to review, the size of the Negative Set is typically much larger than that of the Positive Set, so simply comparing percentages, without translating those percentages into numbers of

[18] On the terms *false negative*, *true positive*, *true negative*, and *false positive*, see the glossary in Chapter 3.

[19] To illustrate with a simple example, suppose, in evaluating a set of search terms, our sampling from the Negative Set had provided us with a false-negative estimate of 1,000 documents. That number, in isolation, provides us with insufficient information to evaluate the effectiveness of the search terms, for there are circumstances in which missing 1,000 responsive documents would be indicative of ineffective search terms and circumstances in which missing 1,000 responsive documents would be consistent with effective search terms. Suppose, however, that we had coupled our sampling from the Negative Set with sampling from the Positive Set and that the latter sampling had provided us with a true-positive estimate of 10,000 documents. In that circumstance, we have enough information to evaluate the effectiveness of the search terms and, more specifically, enough information to find (pending results of a qualitative analysis) that the search terms were performing reasonably effectively (capturing 10 responsive documents for every 1 responsive document missed, a solid result (*pending, it is worth repeating, the qualitative evaluation of the observed false negatives*)). Suppose, on the other hand, that our sampling from the Positive Set had provided us with a true-positive estimate of 2,000 documents. In that circumstance, we would have enough information to find that the search terms were performing poorly (missing one responsive document for every two captured, a poor result).

documents can be misleading.[20] If we want a clear view of the effectiveness of the search terms, we must compare the estimate of the *number of responsive documents* captured to that of the *number of responsive documents* missed.

Why not a recall number? A third question that a practitioner might ask is: *When you say we should compare the number of responsive documents to be included in the downstream review to the number of responsive documents to be excluded from further review, why don't you just go ahead and say that we should calculate the recall achieved by the search terms? After all, if we have estimates of the number of responsive documents included and the number of documents excluded, we have all the inputs we need to arrive at a recall estimate (where* ***recall = included / (included + excluded))****.*

The answer does not have to do with the technical conditions for calculating recall; if we carry out the validation exercise as specified in the Protocol, we will indeed have all the inputs required to arrive at an estimate (and associated margin of error) of the recall realized by taking the exclusionary step under evaluation.[21] Parties (both requesting and responding) may make those calculations if they wish.[22] The reason the Protocol does not make the calculation of the recall achieved by an exclusionary step a *requirement* is a practical one. At the stage in which exclusionary steps are applied to collected data, the prevalence of responsive documents is generally very low (especially if collection has been thorough and broad). When prevalence is very low, the margins of error associated with recall estimates are, even with quite large sample sizes, large. As a result, if the Protocol made the calculation of recall estimates (and associated margins of error) a requirement, it could induce two negative responses on the part of the parties to a discovery effort. First, it could lead the parties down a contentious, and ultimately unproductive, path of trying to reduce the recall margins of error by increasing sample sizes to unmanageable levels. Second,

[20] A failure to convert percentages into counts of documents was likely the reason for an incorrect assessment of the data in a ruling in the *Biomet* case of 2013 (*In re: Biomet M2a Magnum Hip Implant Products Liability Litigation*, MDL 2391, Cause No. 3:12-MD-2391 (N.D. Ind., South Bend Division, Apr. 18, 2013)). In that ruling, a judge, having been asked to assess the impact of an exclusion based on search terms, and having been told that sampling of the Negative Set found that the percentage of responsive documents in that set was between 0.55% and 1.33%, concluded that "a comparatively modest number of [additional responsive] documents would be found" by further review of the Negative Set (p. 5). Conversion of the percentages into counts of documents, however, shows that the number of responsive documents recoverable from further review of the Negative Set was in fact large (85,800 to 207,480), both in absolute terms and when compared to the estimated number of responsive documents in the combined Positive and Negative Sets (267,150 to 481,650). Because the impact of the exclusion, in the defendant's memo to the judge, was quantified in percentages, the judge did not appreciate the real impact of the exclusion.

[21] And practitioners wishing to do so can follow the steps described in Section 1.3 (*On the Validation of a Review Process*). It is also worth noting that the numbers generated by the validation exercise for exclusionary steps may be used, in conjunction with the numbers generated by the validation exercise for the review process, to obtain an estimate of "end-to-end" recall (i.e., recall that reflects the effects of both the exclusionary steps and the review process). Arriving at such a recall estimate would simply be a matter of using the numbers generated by the two validation exercises (more specifically, the outcome of the review of the Negative Sample(s) used for validating exclusionary steps and the outcome of the review of both the Positive and Negative Samples used for validating the review process) as inputs to a stratified estimate of aggregate recall; see Section 1.4 (*Additional circumstances and metrics*). For an instance in which the Court required an estimate of recall that covered both search-term culling and TAR, see *In re Diisocyanates Antitrust Litig*., MDL No. 2862, 2021 WL 4295729 (W.D. Pa. Aug. 23, 2021). The model protocol that is the focus of these guidelines does not, for the practical reasons noted in this section, make the calculation of end-to-end recall a requirement, but it does provide for the generation of the inputs needed for such a calculation (and parties are free to make that calculation if they wish).

[22] If the disclosure provisions the Protocol are adhered to, both requesting and responding parties will have all the inputs needed for the calculations.

it could lead the responding party to narrow the scope of its collection efforts in an effort to realize a higher prevalence in the collected data (in effect, applying a more restrictive filter at the collection stage), a narrowing that could result in a loss of important responsive information at a stage at which there are little or no empirical checks on the possibility of such loss. In order to avoid inducing these adverse behaviors, the Protocol's default requirement, for evaluating exclusionary steps, is simply to calculate estimates for the total number of responsive documents in both the Positive and Negative Sets and to compare those numbers.

How to evaluate the numbers? A final question that a practitioner might ask has to do with evaluating the results: *Once we have done the calculations and have our estimates (and margins of error), how are we to make meaningful use of them? What numbers are indicative of an empirically well-grounded exclusionary step and what numbers indicative of one that is inadvisable?*

The answer is that there is no consensus on a single number (or set of numbers) that would be applicable in all circumstances. We can, however, specify a threshold that, *in most circumstances*, will be both practically realizable and indicative of a reasonably effective result. That threshold is ***a ratio of 10 included responsive documents to every 1 excluded responsive document***; put another way, the point estimate for the number of responsive documents that reside in the Positive Set should be ten times that for the number of responsive documents that reside in the Negative Set.[23]

For example, in terms of the effectiveness of search terms, a ratio of 10 included to 1 excluded corresponds to 90.9% recall. Such a result would be indicative, pending the results of qualitative analysis, of a highly effective retrieval effort (as it should be at the stage at which search terms are applied). In terms of practicality, achieving an included-to-excluded ratio of 10-to-1 will in most circumstances be feasible. Precision is not (or should not) be a primary goal at the stage at which search terms are being applied, so allowing low levels of precision (i.e., allowing the search terms to be broad) should allow practitioners to arrive at search terms that achieve the target ratio (even if that comes at the expense of having to include a good number of non-responsive documents in the Review Set as well).

Finally, with regard to evaluating the numbers, it should be remembered that any quantitative result must be supplemented with information provided by a qualitative analysis of any missed responsive documents uncovered via the validation exercise. We can tolerate missing large numbers of responsive documents that contain unimportant or redundant information, but missing even small numbers of responsive documents that contain important and unique information may require modifying the criteria being applied to make the exclusion (e.g., broadening the search terms) or skipping the exclusionary step altogether.

Section 1.3: On the Validation of a Review Process

The primary question we seek to answer when validating the results of a review for responsiveness (whether that review is manual or some variety of technology-assisted) is how effective the review has been at identifying the responsive documents that reside in the Review Set. The metric that most directly answers that question is *recall*. Recall tells us, out of all the responsive documents that reside in the Review Set,

[23] It is of course always necessary to take into consideration the *nature* of any responsive documents that are missed. The quantitative threshold specified here is meaningful only if supplemented by qualitative analysis that shows that such responsive documents as are excluded are neither important nor novel.

what percentage the review process successfully identified. High recall is generally indicative of an effective review process;[24] low recall is indicative of an ineffective process.

The calculations required to obtain point estimates (and associated margins of error) for recall require a few steps beyond those required to estimate a proportion or total in a single population but are still simply a matter of applying the appropriate equations from the *Equation Library*. The specifics are as follows.

Inputs

The calculation of recall requires that we first obtain estimates of the total number of responsive documents in two distinct populations (the *Positive Set*: the set of documents designated by the review process as responsive (along with any associated family members) and the *Negative Set*: the set of documents neither designated as responsive nor associated with a document designated as responsive). We then combine those two estimates to arrive at the summary metric that is recall. This means that a total of six inputs are required to obtain a recall estimate (and associated margin of error): three to obtain an estimate of the total number of responsive documents in the Positive Set and three to obtain the estimate of the total number of responsive documents in the Negative Set. These inputs are the following:

- N_+: The number of documents in the Positive Set;
- n_+: The number of documents in the Positive Sample;
- r_+: The number of responsive documents observed in the Positive Sample;
- $N_\circ$: The number of documents in the Negative Set;
- $n_\circ$: The number of documents in the Negative Sample; and
- $r_\circ$: The number of responsive documents observed in the Negative Sample.

With regard to N_+ and $N_\circ$, it may be observed that these numbers can be obtained once the review process[25] has been completed (completed, that is, pending the results from the validation exercise). The Positive Set is defined as the set of documents designated as responsive by the review process, together with any associated family members of such documents. The Negative Set is defined as the remainder of documents in the Review Set (i.e., the set of documents neither designated as responsive by the review process nor associated, by family relation, with a document so designated).

With regard to n_+ and $n_\circ$, it may be observed that default specifications for these values (the sample sizes) are given in the Protocol (400 for n_+ and 3,400 for $n_\circ$). As stated in the Protocol, practitioners may depart from the default specifications when circumstances warrant; when Practitioners do so, n_+ and $n_\circ$ will of course represent the sizes of the samples actually drawn.

[24] Many circumstances will require, in the interest of the usability of the results, high recall coupled with at least reasonably high *precision* (for more on precision, see the glossary entry). In terms of meeting the primary requirement of identifying and producing the documents responsive to a production request, however, recall is the salient metric and high recall is the goal (hence the centrality of recall in validating review processes).

[25] It may be worth emphasizing here that the *review process* is defined as the aggregate of all review processes applied to the Review Set. The Positive Set will therefore contain any documents (along with associated family members) designated as responsive by any of the processes (manual or technology-assisted) brought to bear in the review for responsiveness. The Negative Set will be similarly defined based on the aggregate results of all review processes brought to bear in the review for responsiveness.

With regard to r_+ and $r_\circ$, it may be observed that, of the six input numbers, it is only r_+ and $r_\circ$ that require additional document review. The validation samples must be drawn and manually reviewed in order to obtain the required numbers.

Procedures

Once the input numbers are in hand, obtaining an estimate and margin of error for recall is a matter of executing the following steps.[26]

1) Obtain **point estimates** of the **total** number of responsive documents in both the **Positive Set** and in the **Negative Set**.
   - a) Find the point estimate for the number of responsive documents in the Positive Set (t_+).
     - i) Using as inputs n_+ and r_+, apply Equation 1 to obtain the estimated proportion of responsive documents in the Positive Set (p_+).
     - ii) Using as inputs p_+ and N_+, apply Equation 3 to obtain the point estimate for the number of responsive documents in the Positive Set (t_+)
   - b) Find the point estimate for the number of responsive documents in the Negative Set ($t_\circ$): repeat the steps specified under 1(a), replacing the Positive-Set inputs with the corresponding Negative-Set inputs ($N_\circ$, $n_\circ$, $r_\circ$).
2) Obtain the **variances** associated with the **total** estimates.
   - a) Find the variance associated with the t_+ estimate ($var(t_+)$).
     - i) Using as inputs N_+, n_+, and p_+, apply Equation 2 to obtain the estimated variance of the proportion estimator ($var(p_+)$).
     - ii) Using as inputs $var(p_+)$ and N_+, apply Equation 4 to obtain the variance of the total estimator ($var(t_+)$).
   - b) Find the variance associated with $t_\circ$ estimate ($var(t_\circ)$): repeat the steps specified under 2(a), replacing the Positive-Set values with the corresponding Negative-Set values.
3) Obtain the **point estimate** for the **recall** achieved by the review process.
   - a) Using as inputs the total estimates for both the Positive Set and the Negative Set (t_+, $t_\circ$), apply Equation 9 to obtain the point estimate for the recall achieved by the review process ($Recall$).
4) Obtain the **margin of error** associated with the **recall** estimate.
   - a) Using as inputs the total estimates for both the Positive Set and the Negative Set (t_+, $t_\circ$) and their associated variances ($var(t_+)$, $var(t_\circ)$), apply Equation 10 to obtain the estimated variance of the recall estimator ($var(Recall)$).

[26] Again, as noted earlier, while the calculations are reasonably simple, some practitioners may still find it helpful, or reassuring, to engage a consultant with the appropriate expertise to provide support in carrying out this part of the validation exercise.

b) Using as input $var(Recall)$, apply Equation 5 to obtain the margin of error associated with the recall estimate ($M(Recall)$).

5) Converting proportions to percentages,[27] summarize the result.

a) Recall = $Recall \pm M(Recall)$.

Interpreting the numbers

The numbers generated by the validation exercise are, of course, important. They are not, however, ends in and of themselves; they offer one perspective (an essential empirical perspective, to be sure, but still just one perspective) on the question of whether the results of the review process represent a reasonably complete response to the production requests. This section provides some further guidance on interpreting the numbers.

No universally applicable threshold. As observed in the Commentary to the Protocol,[28] there is no consensus around a specific minimum value for recall that a review effort must meet to qualify as "reasonably complete." That is as it should be: there are simply too many circumstance-specific variables that affect both what should be considered "reasonable" and what should be considered "complete" to arrive at a single number that will be appropriate in all circumstances. As a practical matter, however, it is useful for practitioners to have a *prima facie* threshold that can serve as an actionable target until specific circumstances suggest otherwise. The threshold pegged in the Protocol, 75% recall, serves this purpose well, on grounds of both reasonableness and completeness.

75% recall is an achievable target. The achievement of 75% recall is a challenge, but one that can be met with the competent operation of advanced review technologies. As observed in the Commentary to the Protocol, in the studies conducted in the TREC Legal Track,[29] of the 53 submissions to the Interactive Task from 2008 to 2010, seven were found to have achieved recall of 75% or greater; and, of those seven, five achieved that threshold while also meeting or exceeding 75% precision. Of the 70 submissions to the Track's Learning Task in 2011, none met the 75% recall threshold while also maintaining at least 75% precision (or even while also maintaining at least 50% precision), but several[30] were able to meet the 75% recall threshold at lower levels of precision. Exercises conducted in the TREC Total Recall Track in 2015 and 2016 provide further evidence that 75% recall is an achievable threshold.[31] From the 2015 edition of

[27] Converting a proportion to a percentage is simply a matter of multiplying the proportion by 100 (and adding the % symbol to the expression); thus, for example, the proportion 0.05 converts to 100 x 0.05 = 5%. Both proportions and percentages are suitable modes for expressing recall; we convert proportions to percentages in summarizing the results simply because that is currently the most common practice in the field of e-discovery.

[28] See the Comment on *A reasonable prima facie threshold*.

[29] For a complete archive of resources related to the TREC Legal Track, see: https://trec-legal.umiacs.umd.edu/. For summaries of the results of each edition of the Legal Track, see the track overviews for each year (available at the resource page just noted and at: https://trec.nist.gov/proceedings/proceedings.html).

[30] To be specific, using the cutoffs reported in the Track Overview for 2011, 19 submissions achieved recall of 75% or greater while also maintaining precision of at least 2%. For the Learning Task, results were reported at different "cutoff" points down the ranked lists of documents submitted by task participants. The achievement of any target level of recall (even 100%) is possible as long as you choose a cutoff deep enough down the list (although that will often coincide with a very low level of precision). (A minimum threshold for precision of 2% is quite low; whether that level of precision would be acceptable in return for recall of 75% or higher would be dependent on the specific circumstances that occasioned the review exercise.)

[31] Grossman et al. 2016; Roegiest et al. 2015.

the track, for example, we find that, if we consider each participant's best run for each of the five test collections featured in the exercise, over half of the runs (21 out of 34) achieved, averaging across the topics featured in each collection, recall of at least 75% while maintaining precision of at least 50%.[32] Asking that a review exercise should meet a minimum threshold of 75% recall is not asking the impossible or even the unreasonable, especially when we allow that a responding party has the option of accepting lower levels of precision if need be.[33]

Redundancy of information across documents means that information gain may decline at higher levels of recall. When thinking about numbers, it is helpful to keep in mind the redundancy that characterizes the distribution of information across the documents in a collection. In the collections of documents typically subject to legal discovery, the information salient to discovery requests is not distributed such that each document contains its own unique bit of information; rather the information is typically distributed in a redundant (one-to-many) fashion, such that the same information is often contained in many documents.

What this means is that, on the one hand, it is true that there is a general correlation between retrieving documents and retrieving information: generally speaking, the greater the number of responsive documents we have retrieved, the greater the amount of salient information we will have retrieved. Hence the value of the (document-based) recall metric as a measure of the effectiveness of a review. On the other hand, the correlation is not perfect. Given the redundancy of information, the return on investment (new information gained from documents newly retrieved) may decline as the number of documents already retrieved increases. In most cases, therefore, once a review has achieved a reasonably high level of recall (e.g., 75%), the unretrieved (missed) documents will add little new or important information to that which can already be gathered from those that were successfully retrieved.[34] A threshold of 75% recall is thus normally suitably high to serve as a reasonable *prima facie* threshold for completeness. The "*prima facie*" qualification is important here, however: the way in which information is distributed across documents will vary from one review set to the next, so the validity of the hypothesis that, at a given level of recall, any missed documents will add little novel or important information must be corroborated, for the specific review set at hand, by a qualitative analysis of observed false negatives (missed responsive documents).

A small margin of error is not the goal of a validation exercise. It is important to be transparent about the margin of error associated with a statistical estimate; that is how one gains a sense of the scope of uncertainty associated with the estimate. That said, it is also important to remember that a small margin of error is not the goal of the validation exercise. The goal of the validation exercise is to provide empirical evidence that will enable an assessment as to whether the results of a review process are reasonably complete. That goal can sometimes be met even with rather large margins of error. Practitioners' meet-and-confer discussions about the design of a validation exercise, and about the results of one, will be more productive if the fundamental goal of validation is kept in mind and if the discussions are not allowed to be diverted down unproductive paths about whether the margin of error should meet an arbitrary pre-specified limit. Often what can and cannot be accomplished (with regard to reducing the margin of error), while

[32] Roegiest et al. 2015.

[33] This discussion of thresholds assumes, of course, that the recall estimate in question has been obtained by sound methods (an appropriate sampling design, blind review of samples, proper execution of estimation procedures), such as those described in these guidelines and the protocol they are intended to support.

[34] For additional discussion of the distinction between document recall and information recall, see Hedin et al. 2016: 412.

remaining within manageable (and proportionate) limits on sample size, cannot be known until we are at the stage of actually conducting the validation exercise.

Remember the qualitative perspective. Quibbles about numbers are generally best addressed by qualitative analysis. Any quantitative measure must be supplemented by qualitative analysis. It is not impossible that, when a review has achieved 75% recall, there is still a significant number of important documents yet to be retrieved.[35] Conversely, it may be the case that, when a review has achieved 70% recall, it has indeed captured, in some form or other, all the genuinely important documents. This is where qualitative assessment comes in; it enables the parties to evaluate the nature of any responsive documents that have been missed by the review process and, more specifically, the nature of the information contained in such documents and whether that information is both genuinely important to the issues being litigated and non-redundant with information that can be gathered from the set of documents the review has successfully identified as responsive. Practitioners will reach agreement on the results of validation more effectively by this sort of analysis than by quibbling about whether recall should be a few percentage points higher.

Section 1.4: Additional Circumstances and Metrics

Our discussion so far has focused on the simplest sampling scenario (the scenario in which the review has generated just one Positive Set and one Negative Set) and on the metrics of primary interest in a validation exercise (recall or, in the case of the validation of an exclusionary step, the total number of responsive documents that reside in a single population). While, however, the simple review scenario is indeed common, there are other review scenarios, such as when an incremental or rolling review and production generates multiple Positive or Negative Sets, that are not uncommon. And while recall is the most telling metric in validating a review process, there are other metrics, such as precision and prevalence, that can provide valuable color and context for evaluating the significance of the main metrics. In this section, we outline the steps to be followed when validation goes beyond the "canonical" case. We focus, more specifically, on estimation when circumstances require a stratified design and on the estimation of the ancillary metrics *precision* and *prevalence*.

Circumstances requiring a stratified design

It is not uncommon, especially when the collection of documents potentially subject to review extends over a long period of time or when the review itself does so, or when the scope of the review effort expands, that one or more of the sets that are in-scope for a validation exercise are not simple but compound. That is to say that a set that is a domain for estimation is in fact made up of multiple distinct component subsets (rather than being one undifferentiated set). In the case of collection, for example, an initial collection effort may focus on readily accessible data sources and the results of that effort may then be filtered by search terms. A subsequent collection effort may focus on less immediately accessible data sources. When the results of the latter effort are also filtered by search terms, the result is multiple Positive and Negative Sets (one each from the initial collection effort and one each from the subsequent effort). In the case of a review for responsiveness, the responding party may choose to meet its discovery obligations by making "incremental" or "rolling" productions from the data subject to review, resulting in multiple Positive Sets.

[35] For further discussion of the implications of this scenario, see Grossman & Cormack 2021: 26.

When such compound sets are in-scope for validation, the estimation procedures required are the same as those followed for simple sets, with the exception that an initial aggregation step must be taken in order to arrive at overall numbers for the compound set.[36] Those aggregate numbers (estimates and variances) then get plugged into the same equations used to obtain the relevant metrics for simple populations.[37] To illustrate, we look, first, at estimating an aggregate total from multiple strata (as we do in validation of search terms) and, second, at estimating aggregate recall with multiple Positive and Negative Sets.

Estimating an aggregate total

As we have seen, in the case of a simple population, arriving at an estimate of the total number of responsive documents in the population (and the variance associated with that estimate), is a matter of applying Equations 1 through 4. In the case of a compound population (i.e., a population composed of multiple subsets or *strata*[38]), obtaining an aggregate estimate (and associated variance) of the total number of items of interest in the full population is a matter of, first, applying the same equations (1 through 4) to obtain estimates and variances for each stratum that is a component of the full population and then applying Equations 7 and 8 to aggregate the stratum-specific numbers to full-population results. More specifically, we proceed as follows.[39]

1) For each stratum, obtain **the stratum-specific point estimate** and the **stratum-specific variance** associated with that estimate ($t_{(i)}$, $var(t_{(i)})$.
    a) Using the stratum-specific inputs $N_{(i)}$, $n_{(i)}$, and $r_{(i)}$, apply Equation 1 and Equation 3 to obtain the point estimate for the number of responsive documents in each stratum ($t_{(i)}$).
    b) Using the stratum-specific inputs $N_{(i)}$, $n_{(i)}$, and $p_{(i)}$, apply Equation 2 and Equation 4 to obtain the estimated variance of the total estimator for each stratum ($var\left(t_{(i)}\right)$).
2) Obtain the **aggregate total estimate**: using as input the stratum-specific total estimates ($t_{(i)}$ for $i = 1$ to L), apply Equation 7 to obtain the full-population estimate of the number of responsive documents ($t_{(ag)}$).
3) Obtain the **aggregate variance** associated with the full-population estimate: using as input the stratum-specific variances ($var\left(t_{(i)}\right)$ for $i = 1$ to L), apply Equation 8 to obtain the estimated variance of the full-population total estimator ($var\left(t_{(ag)}\right)$).
4) Obtain the **aggregate margin of error** associated with the full-population estimate: using as input $var\left(t_{(ag)}\right)$ (the output of the preceding step), apply Equation 5 to obtain the margin of error

[36] And that aggregation step is, following procedures drawn from stratified sampling, largely a matter of summing the individual results obtained on specific subsets.
[37] Alternatively, a party may choose simply to wait until all the in-scope subsets have been reviewed and only then conduct the validation exercise (which could then follow the procedures for the simple "canonical" case).
[38] For the term *stratum* (pl. *strata*), see the entry for *Stratified Sampling* in Chapter 3.
[39] In the following, we assume a case in which our goal is to estimate the aggregate total for a full population, without any differentiation among positive or negative sets (i.e., our goal is to arrive at the aggregate estimate $t_{(ag)}$ from the stratum-specific estimates $t_{(i)}$ for strata $i = 1$ to $i = L$). The same steps would hold, however, with appropriate changes to the inputs, for obtaining aggregate total estimate for subsets of the full population (e.g., $t_{+(ag)}$, $t_{\circ(ag)}$).

associated with the full-population estimate of the total number of responsive documents ($M\left(t_{(ag)}\right)$).

5) **Summarize** the result. Total responsive documents in the population: $t_{(ag)} \pm M\left(t_{(ag)}\right)$.

Estimating aggregate recall

Circumstances requiring the estimation of aggregate recall using stratified designs are not uncommon. The steps to follow in these circumstances are as follows.[40]

1) Following the steps provided in the preceding section (*Estimating an aggregate total*), obtain the **aggregate point estimates and variances** used in the calculation of recall: $t_{+(ag)}$, $var\left(t_{+(ag)}\right)$, $t_{\circ(ag)}$, and $var\left(t_{\circ(ag)}\right)$.
2) Using the output of the preceding step, apply Equation 9, Equation 10, and Equation 5 to obtain the **point estimate** for recall and the **margin of error** associated with that estimate ($Recall_{(ag)}$, $M\left(Recall_{(ag)}\right)$.
3) **Summarize** the result. Aggregate recall: $Recall_{(ag)} \pm M\left(Recall_{(ag)}\right)$.

Additional metrics

While the primary focus of an evaluation of the effectiveness of a review process is generally recall, there are other metrics (namely, precision and prevalence) that can provide valuable context for evaluating the results of recall estimation (and for assessing the resources required to arrive at recall estimates). In the following we review the steps required to obtain estimates of these metrics, starting with precision and then turning to prevalence.[41]

Precision

1) Following the steps provided in the section *Estimating an aggregate total*, obtain the **aggregate point estimates and variances** used in the calculation of precision: $t_{+(ag)}$ and $var\left(t_{+(ag)}\right)$.
2) Using the sizes ($N_{+(i)}$) for each of the i to L positive strata, apply Equation 6 to obtain the **population size** ($N_{+(ag)}$) relevant for the calculation of aggregate precision.
3) Using the output of the preceding steps, apply Equation 11, Equation 12, and Equation 5 to obtain the **point estimate** for precision and the **margin of error** associated with that estimate ($Precision_{(ag)}$, $M\left(Precision_{(ag)}\right)$.

[40] In the steps provided here, we assume that both the Positive Set and the Negative Set are composed of multiple strata and so the relevant estimates from both sets require aggregation via procedures applicable to stratified sampling designs. It is not always the case, however, that both sets are compound. In some cases, for example, there may be multiple positive subsets (strata) but just one Negative Set. In those cases, we apply the stratified procedures to the compound (stratified) set but use the simpler procedures given earlier for the non-stratified set.

[41] For both metrics we provide the steps required in the more complex case in which stratified estimation is required. When circumstances do not require stratified estimation, we can use the simpler non-compound population inputs to the calculations.

4) **Summarize** the result. Aggregate precision: $Precision_{(ag)} \pm M\big(Precision_{(ag)}\big)$.

Prevalence

1) Following the steps provided in the section *Estimating an aggregate total*, obtain the **aggregate point estimates and variances** used in the calculation of prevalence: $t_{(ag)}$ and $var\big(t_{(ag)}\big)$. (Note that, in the calculation of prevalence, which tells us the percentage of responsive documents in the full population at issue in a validation exercise, we sum across *all strata* (both positive and negative) to arrive at the aggregate numbers.)
2) Using the sizes ($N_{(i)}$) for each of the i to L strata in the population for which prevalence is being estimated, apply Equation 6 to obtain the **population size** ($N_{(ag)}$) relevant for the calculation of aggregate prevalence.
3) Using the output of the preceding steps, apply Equation 13, Equation 14, and Equation 5 to obtain the **point estimate** for prevalence and the **margin of error** associated with that estimate ($Prevalence_{(ag)}$, $M\big(Prevalence_{(ag)}\big)$.
4) **Summarize** the result. Aggregate prevalence: $Prevalence_{(ag)} \pm M\big(Prevalence_{(ag)}\big)$.

Section 1.5: Worked Examples

In Appendix B, we walk through two examples of applying the procedures for validating an exclusion based on search terms and two examples of applying the procedures for validating a review for responsiveness. Readers are encouraged to strengthen their familiarity with the procedures by working through these examples.

Chapter 2: A Guide to Sample Size Selection

Among the more contentious topics raised in meet-and-confer discussions is that of sample size: *How large should the samples be that are used in a given validation exercise?* The requesting party typically seeks larger sizes; the responding party argues for smaller. The contentiousness of such discussions can be reduced by basing the discussion on a clear-eyed view of what can be accomplished by samples of different sizes. In this chapter, we examine the question of the power of various sample sizes and show the evidence and reasoning that are the basis for the default sample sizes specified in the Protocol.

One initial note on sample size is in order before we turn to specifics. In the discussion of sample sizes that follows, we will be discussing the properties of different sample sizes and, more specifically, the power of different sample sizes to contain the sampling error that contributes uncertainty to our estimates. In assessing these properties, it is important to remember that, as long as the sample size is small relative to the size of the population from which the sample is drawn (as it almost always is in the case of the sorts of validation exercises we are considering), the specific size of the population does not affect the power of a given sample size to contain sampling error**: the capabilities described for a given sample size hold regardless of the size of the population from which the sample is drawn**.[42] We therefore can consider the capabilities of different sample sizes and arrive at recommendations for sample size independent of population size.

Section 2.1: On the Size of Samples Other than the Recall Negative Sample

As provided for in the Protocol, a number of samples may need to be drawn and reviewed for validation purposes in the course of responding to a request for production: the samples used to validate the exclusion of data from the Review Set, for example, or, when validating the results of the review itself, the samples used to obtain an estimate of recall. The primary focus of this chapter is on the size of the *Negative Sample* required to obtain precise estimates of recall. We place the focus here because the size of the Negative Sample is the key factor[43] in determining how well we are able to control for the sampling error that contributes uncertainty to our recall estimates (more concretely, how much we are able to constrain the margins of error associated with our recall estimates) and because discussions around the size of the Negative Sample are, given the compound nature of the recall estimate and the resulting complexity in obtaining a view of the impact of variations in sample size, most in need of the guidance provided by an empirical overview of the power of different sample sizes.

This is not to say that questions about the sizes of other samples used for validation are not important. They are important, but they are also more easily addressed.

[42] This derives from the fact that, once a sample is small relative to the population from which it is drawn, sampling error, the uncertainty that derives from the random selection process, is, for all practical purposes, entirely a function of the size of sample; the size of the population ceases to be a factor. As stated in Freedman et al. 2011: 367: "When estimating percentages, it is the absolute size of the sample which determines accuracy, not the size relative to the population. This is true if the sample is only a small part of the population, which is the usual case." Also see Thompson 2002: 36.

[43] More precisely, the key factor that is within our control. There are other factors, the most important of which are the overall prevalence of responsive documents in the Review Set and the level of recall actually achieved by the review process. These factors, however, are not in our control at the point of designing and executing a validation exercise.

Samples used in the validation of exclusionary steps. More specifically, with regard to the sizes of samples used to validate the exclusion of data from the Review Set (whether that exclusion is based on metadata values, the results of applying search terms to the Collected Set, or on some other basis), we note that the Protocol specifies a size of 400 documents for the Positive Sample and, for the Negative Sample, a size of 6,000 documents (when validating an exclusion based on search terms) or 1,200 documents (when validating a metadata-based exclusion). With regard to these sizes, we can say the following.

- **Positive Sample: 400 documents.** The purpose of the Positive Sample is to obtain an estimate of the number of responsive documents in the Positive Set (the set that will be included in the downstream review); this estimate will be used as a reference against which to compare the analogous estimate obtained from the Negative Sample (i.e., the number of responsive documents in the Negative Set). A sample of 400 documents will have the following characteristics.
  - It will enable us, in gauging the uncertainty associated with our estimate, to obtain a margin of error[44] that is no greater than ± 5%.[45]
  - It will almost always[46] include at least one instance of any type of document that is represented in at least 0.75% of the source population (i.e., occurs at a frequency of at least 1 out of every 133 documents).
  - It will, in most cases,[47] bring into view a meaningful range of the sorts of responsive documents included in the Review Set.
- **Negative Sample (search-term culling): 6,000 documents.** The purpose of the Negative Sample is to obtain an estimate of the number of responsive documents in the Negative Set (the set that will be excluded from downstream review); this estimate will be compared against the analogous estimate obtained from the Positive Sample in order to assess the quantitative impact of the exclusion. A sample of 6,000 documents will have the following characteristics.
  - It will enable us, in gauging the uncertainty associated with our estimate, to obtain a margin of error[48] that is no greater than ± 1.3%.
  - It will almost always[49] include at least one instance of any type of document that is represented in at least 0.05% of the source population (i.e., occurs at a frequency of at least 1 out of every 2,000 documents).
  - It will allow a view of the sorts of responsive documents that would be excluded from the Review Set (assuming that such documents exist).

[44] Calculated at a level of 95% statistical confidence.

[45] To express the size of the margin of error in a manner that allows easy comparison from one circumstance to another, we assume the estimate of responsive documents in the Positive Set is expressed as a percentage (rather than an absolute number). In any specific circumstance, practitioners would be advised also to express both the estimate and the margin of error in terms of numbers of documents (see Chapter 1, under the heading *Why a number, not a percentage?*).

[46] To be specific: 95% of the time.

[47] The exceptions being cases in which the prevalence of responsive material in the set designated for inclusion in the Review Set is still very low. Such cases, when they occur, can be addressed by either increasing the size of the Positive Sample or re-visiting the exclusionary technique in question (e.g., search terms) to make the technique more effective at avoiding the inclusion of false positives.

[48] Calculated at a level of 95% statistical confidence.

[49] To be specific: 95% of the time.

- **Negative Sample (metadata-based exclusion): 1,200 documents.** The purpose of the Negative Sample is to obtain an estimate of the number of responsive documents in the Negative Set (the set that will be excluded from downstream review); this estimate will be compared against the analogous estimate obtained from the Positive Sample in order to assess the quantitative impact of the exclusion. A sample of 1,200 documents will have the following characteristics.
  - It will enable us, in gauging the uncertainty associated with our estimate, to obtain a margin of error[50] that is no greater than ± 2.8%.
  - It will almost always[51] include at least one instance of any type of document that is represented in at least 0.25% of the source population (i.e., occurs at a frequency of at least 1 out of every 400 documents).
  - It will allow a view of the sorts of responsive documents that would be excluded from the Review Set (assuming that such documents exist).

Samples with these characteristics should, in most circumstances, provide sufficiently precise estimates of the target metrics to enable a meaningful assessment of the impact of the exclusion being evaluated.[52] As a practical matter, the cost of reviewing samples of the specified sizes, while not insignificant, is not unduly burdensome, given the potential impact of excluding data from further review. If the cost of reviewing samples of the specified size outweighs the savings that would be realized by taking a given exclusionary step (such as using search terms to define the Review Set), then it is best to skip the exclusionary step and include the data in the downstream review. If the savings that would be realized by taking a given exclusionary step do outweigh the cost of reviewing the samples, then taking the exclusionary step is a reasonable means for improving the efficiency of the review (assuming, of course, that the results of the validation exercise support the exclusion).

The Positive Sample used in the estimation of recall. When we estimate recall, we draw two samples: the Positive Sample (the sample drawn from the set of documents (and associated family members) designated by the review as responsive to the requests for production) and the Negative Sample (the sample drawn from the set of documents neither designated as responsive nor associated, by family relation, with a document designated as responsive). As already noted, the focus of this chapter is on the size of the Negative Sample used in the estimation of recall. With regard to the *Positive Sample*, the Protocol specifies a size of 400 documents; about a Positive Sample of this size, we can say the following.

- It will ensure that the margin of error[53] associated with our estimate of the percentage of responsive documents in the Positive Set is no greater than ± 5%.[54] It should also be noted that, given the

[50] Calculated at a level of 95% statistical confidence.

[51] To be specific: 95% of the time.

[52] Of course, as provided for in the Protocol, quantitative measures need to be supplemented by qualitative analysis to obtain a complete view of the impact of a given exclusion.

[53] Calculated at a level of 95% statistical confidence.

[54] Again, to express the size of the margin of error in a manner that allows easy comparison from one circumstance to another, we assume the estimate of responsive documents in the Positive Set is expressed as a percentage (rather than an absolute number). In any specific circumstance, practitioners would also be advised to express both the estimate and the margin of error in terms of numbers of documents (see Chapter 1, under the heading *Why a number, not a percentage?*).

definition of precision,[55] this sample size will allow us to obtain an estimate of precision that has a margin of error no greater than ± 5%.

- It will sufficiently constrain the uncertainty associated with the estimate of the number of responsive documents that reside in the Positive Set[56] to ensure that the Positive component of the recall estimate is not a practically significant contributor to the margin of error associated with our eventual estimate of recall.
- It will almost always[57] include at least one instance of any type of document that is represented in at least 0.75% of the source population (i.e., occurs at a frequency of at least 1 out of every 133 documents) and will thus bring into view a meaningful range of the sorts of responsive documents successfully identified by the review process as well as of the sorts of non-responsive documents the review process *incorrectly* classified as responsive.

Section 2.2: On the Size of the Recall Negative Sample

Having reviewed the grounds for the Protocol's size specifications for the samples used in validating exclusionary steps and for the Positive Sample used in the estimation of recall, we can now turn, in the remainder of this chapter, to the more challenging question of the size of the Negative Sample used in the estimation of recall. We begin by describing an approach to gauging the power of a given sample size to constrain the sampling error that contributes uncertainty to a recall estimate. Applying that approach, we then arrive at a set of candidate sample sizes[58] for the Negative Sample. We follow with further analysis to confirm the advisability of the candidate sample sizes. In a brief concluding section, we summarize our recommendations for the size of a Negative Sample used in the estimation of recall.

Analyzing the power of a sample

In estimating recall, we have, as we have seen, two samples:[59] a Positive Sample and a Negative Sample. Let us assume, to start a running example, that our review has resulted in a Positive Set of 200,000 documents and a Negative Set of 1,800,000 documents.[60] From the Positive Set we have drawn a Positive

[55] Also see the discussion of the procedures for calculating *precision* in Section 1.4 (*Additional Circumstances and Metrics*), under the heading *Additional metrics*.

[56] I.e., the uncertainty due to *sampling error*; for a definition of sampling error, see the Glossary.

[57] To be specific: 95% of the time.

[58] We use the plural here because, as will be seen, we specify different sample sizes for different levels of prevalence.

[59] In the canonical case in which we have one Positive Set and one Negative Set. The approach outlined here for selecting a sample size would also work for cases in which we had multiple Positive or Negative Sets; the approach requires just that we take as given the sample sizes that are not in question and focus the analysis on the sample size that is in question (as, in this case, the Negative Sample).

[60] For purposes of the analysis described in this chapter, we assume a scenario in which the review resulted in a Positive Set of 200,000 documents and a Negative Set of 1,800,000 documents (and we draw the Positive and Negative Samples from those sets). A margin-of-error analysis, like the one we are conducting, will get slightly different results if we vary the relative sizes of the Positive and Negative Sets (e.g., if we assume a scenario in which the review resulted in a Positive Set of 500,000 documents and a Negative Set of 1,500,000 documents), even while holding recall and prevalence levels constant. Such differences will not be large, however, unless we assume a starting point for the analysis very different from the one we do assume (e.g., a starting point in which the Positive Set is larger than the Negative Set). For purposes of arriving at default sample sizes, therefore, such differences can safely be set aside. The scenario we assume models typical real-world scenarios reasonably well. If practitioners do encounter a

Sample of the Protocol-specified 400 documents. We now wish to evaluate the power of an 800-document Negative Sample. In a review for responsiveness, assessments are binary: a document is either responsive (designated for inclusion in the production set) or non-responsive (not designated for inclusion in the production set). This means that, for any combination of sample sizes, we have a finite number of possible outcomes. Continuing with our example, for the Positive Sample (400 documents), we have a total of 401 possible outcomes (0 responsive out of 400 sampled, 1/400, 2/400, …, 400/400). For the Negative Sample (800 documents), we have 801 possible outcomes (0/800, 1/800, …, 800/800). For the combination of both samples (as we would have when we use both samples in the estimation of recall), we have 401 × 801 = 321,201 possible outcomes. Assessing the power of a given sample size is then simply a matter of calculating the margin of error for each of the outcomes in that finite set and then assessing the distribution of all the margins so calculated. Some margins of error will be small, some large; our question, in assessing the distribution, will be whether, given the range of possible margin-of-error outcomes, the sample size in question constrains sampling error sufficiently to meet our information needs.

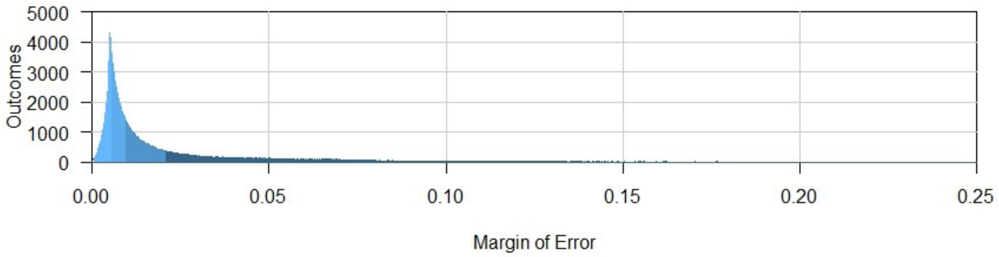
Returning to our example (a Positive Sample of 400 documents; a Negative Sample of 800 documents), we can calculate the margin of error for each of the 321,201 possible outcomes and then assess the distribution of the result. Figure 2.1[61] shows the distribution of margins of error for this sampling design.[62]

circumstance in which the results of the review do vary considerably from those we assume here (in terms of the relative sizes of the Positive and Negative Sets), they can always conduct their own analysis of sampling power (following the method shown in this chapter) assuming appropriately revised inputs. In conducting such an analysis, practitioners may find it helpful to call on the support of individuals with the appropriate statistical expertise (whether these individuals are in-house experts, support specialists provided by a technology vendor, or independent consultants with a practice in legal discovery).

[61] In Figure 2.1, and in other charts like it in this section, we summarize the power of a given sample size by showing, via a ***histogram***, the distribution of the margins of error (for a recall estimate) that result from using that size sample. More specifically, the range of possible margins of error is shown on the horizontal (or "x") axis (in Figure 2.1, the range goes from ± 0% to ± 25%). The height of the blue area (which is actually a series of very narrowly defined bars, each of which is defined for a narrow span on the x-axis) shows the number of outcomes that correspond to a given margin-of-error value: the higher the blue area at a given point on the x-axis (i.e., at a given margin of error), the more outcomes the sample size under evaluation generates at that point (i.e., at that margin of error). To show quartile divisions in the distribution of margins of error, we use blue shading: the lightest blue area represents the first quartile (if we sort the margins of error by size, from largest to smallest, and divide the list into quarters, the lowest quarter (the quarter with the smallest margins of error) would be the first quartile), the next darkest the second quartile, the next darkest the third quartile, and the darkest blue the fourth quartile.

[62] It should be emphasized that these histograms are *not* to be read as probability distributions: they should not be read as implying that, for any given sample size, the more possible outcomes that will generate a given margin of error, the more likely the sample size will, in real-world practice, result in that margin of error. That implication would hold true only if we could say that all possible outcomes for a given sample size were equally likely. We cannot say that, however, because there are other factors (e.g., the level of recall actually achieved by a review process) that may make some outcomes, within the set of possible outcomes for a given sample size, more likely than others. That interpretive boundary observed, the histograms do represent the full set of possible margins of error that could be generated by a given sample size, as well as the number of outcomes consistent with each margin-of-error span (where, again, the number of outcomes consistent with a margin of error ≠ probability of that margin of error), and so are a meaningful and useful gauge of the power of a sample size.

Figure 2.1: Distribution of Margins of Error; No Recall Constraint; No Control on Prevalence



As can be seen from the chart, while there are some outcomes that would result in a large margin of error (there is a long tail to the right, leading to a maximum[63] margin of error of ± 54.5%), most are quite small: the median[64] margin of error is a tiny ± 0.9% and the third quartile[65] value is just ± 2.1%. This result would seem to indicate that the sampling design will suffice to meet our information needs in most cases.

Not so fast, however. If we are to assess the power of the sampling design *where it matters in practice*, we need to introduce controls on the outcomes we include in the analysis to ensure that our view of the power of the sample is not clouded by irrelevant outcomes. More specifically, we need to introduce controls on the levels of recall and the levels of prevalence included in the analysis.

With regard to recall, it may be observed that, of the full set of logically possible outcomes, many (in fact, most) are cases unlikely to occur in practice, simply because the level of recall (as indicated by the point estimate) is so low that a responding party would never submit results in such a state for validation (or, if they did, the point estimate alone, regardless of the size of the margin of error, would suffice to indicate that there were serious problems with the review). To illustrate with our running example, within the set of logically possible outcomes is the case in which we have a 5/400 result for the Positive Sample (5 responsive documents in the 400-document sample) and a 795/800 result for the Negative Sample (795 responsive documents in the 800-document sample). Were we to get this result, our point estimate for recall would be less than 1% (to be exact: 0.14%) and the margin of error associated with that estimate would be ± 0.12%. Such a result, while logically possible, is unlikely to be encountered in practice (and, if it were, it would indicate that the responding party had bigger problems than the size of the margin of error). Including results such as this in the analysis clouds our view of the power of the sample in the cases that matter (the cases we are likely to encounter in practice).

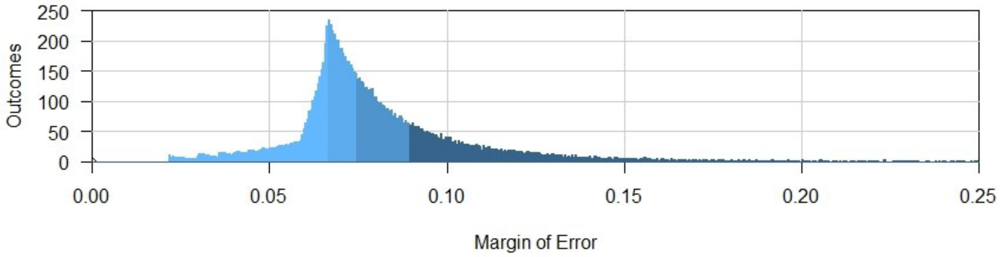
To get a clearer view of the power of the sample, therefore, we limit the analysis to cases which result in a point estimate for recall of 60% or greater (i.e., cases in which recall is at least within shouting distance of being acceptable). Returning to our example, the recall constraint reduces the number of outcomes we include in the analysis from 321,201 to 11,689. The distribution of margins of error for this set of outcomes is represented in Figure 2.2.

[63] The maximum is not represented on the chart; to improve the readability of the chart we have truncated the x-axis at a margin of error value of ± 25%.

[64] The margin of error value that represents a cutoff, in a list of margin-of-error outcomes sorted from largest to smallest, at which 50% of the outcomes would be below the cutoff and 50% above.

[65] The margin of error value that represents a cutoff at which 75% of the outcomes would be below the cutoff and 25% above.

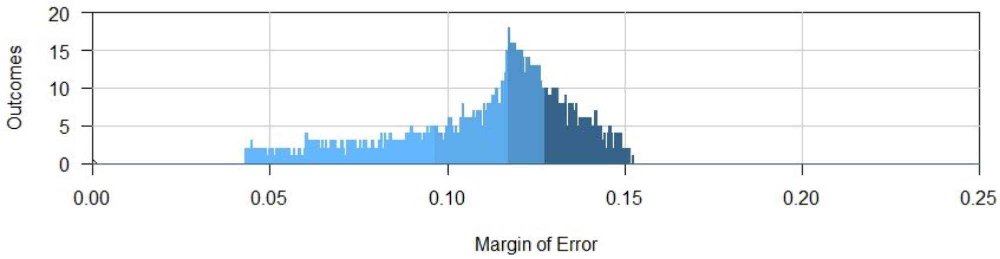
Figure 2.2: Distribution of Margins of Error; Recall Constraint Applied; No Control on Prevalence



As can be seen from the chart, when we focus on just the cases likely to matter in practice, we get a very different view of the power of the sample. The distribution is shifted (relative to that shown in Figure 2.1) considerably to the right: the median margin of error is now ± 7.4%, the third quartile value is ± 9.0%, and the maximum (not shown on the chart) is ± 49.8%. While there may be circumstances in which a sampling design that generated such a distribution of margins of error would be acceptable, there are many other cases in which we would seek a stronger constraint on sampling error (and thus a Negative Sample larger than 800 documents).

Introducing the constraint on recall improves our view of the power of a sampling design, but there is one additional factor for which we need to introduce controls: prevalence. The overall prevalence of responsive material in a Review Set is a strong driver of the size of the margin of error associated with a recall estimate (and hence a strong driver of the sample size required to keep the margin of error within target limits). All other factors being equal, the higher the prevalence, the smaller the margin of error; the lower the prevalence, the larger the margin of error. If, therefore, in assessing the power of a given sample size, we do not apply any controls on prevalence, we will be mixing high-prevalence cases with low-prevalence cases, potentially giving us a muddled view of the power of the sample in the circumstance for which we are conducting the analysis. Returning to our running example, if we restrict our view to cases in which the estimated prevalence would be between 3% and 5% (and continue to apply the recall constraint), we get, as shown in Figure 2.3, a still different view of the power of an 800-document Negative Sample.

Figure 2.3: Distribution of Margins of Error; Recall Constraint Applied; Prevalence between 3% and 5%



As can be seen from the chart, when we restrict our view to cases in which the estimated prevalence is between 3% and 5%, the distribution of margins of error shifts further to the right: the median margin of error is now ± 11.7%, the third quartile value is ± 12.8%, and the maximum is ± 15.2%.[66] What this view

[66] Note that the prevalence constraint we are considering eliminates some of the extremely large margins of error we saw in the cases in which no prevalence controls were applied. It does so because it eliminates, in addition to the high-prevalence cases, cases in which the estimated prevalence is less than 3% (which yield large margins of error).

tells us is that, when prevalence is between 3% and 5%, an 800-document Negative Sample does not constrain the scope for sampling error very well; if, therefore, we think our prevalence is likely to be in that range, we will, unless there are unusual proportionality considerations in play, want to use a larger Negative Sample for our validation exercise.

In order to introduce a control on prevalence, and thus obtain an uncluttered view of the power of a given sample size, we can define distinct prevalence bands and then assess the power of a given sample size for each of those bands separately. In the analysis of sample size that follows, we distinguish among seven prevalence bands, ranging from greater than or equal to 10%, on the high end, to less than 1%, on the low end.

Setting a criterion and finding a sample size

We are now in a position to find candidate sample sizes for the Negative Sample. All that is required is that we specify, in terms of the distribution of margins of error, a criterion that a sample must meet to be acceptable; we then find the smallest sample size that meets that criterion. Given the impact of prevalence on the power of a sample, we specify a distinct criterion for each prevalence band.[67] The results are shown in Table 2.1.

With regard to the prevalence bands, it may be observed that we define the bands more granularly at the lower levels of prevalence. This is done because it is at the lower levels of prevalence that the task of finding manageable sample sizes becomes particularly challenging. Additionally, at the lower levels of prevalence, small differences in prevalence can make big differences in the required sample size. Making finer-grained distinctions in this region of the prevalence space improves our ability to calibrate the required sample size to the circumstance actually at hand.[68]

With regard to the criteria specified for each band, it may be observed that the criteria are progressively loosened as we proceed from higher prevalence bands to lower ones. This is done in recognition of the fact that the sampling error associated with a recall estimate is less easily constrained at lower levels of prevalence and so the criteria must be loosened if we are to arrive at sample sizes that, for low prevalence cases, are at all feasible.[69]

[67] And, to keep the focus on meaningful cases, we consider only outcomes that would yield a recall point estimate of 60% or greater.

[68] It may be worth noting, in terms of the table's structure, that, given the finer grained definition at lower levels of prevalence, the middle band represented in the table is that for prevalence between 3% and 5%. This middle position should not be taken to imply, however, that this band represents the typical case. The prevalence in any given circumstance is the result of many factors (including the methods used to define the Review Set). In real-world practice, cases in the 5% to 10% range of prevalence (i.e., the second and third bands represented in the table) are quite common, as are cases with lower prevalence, particularly when search-term culling has not been employed.

[69] For example, at the second lowest prevalence band (1% to 2%), even a Negative Sample of 200,000 documents will still fall far short of the criteria set for the higher prevalence bands: in the 1% to 2% prevalence range, only 60% of the margins of error generated by a sample of 200,000 documents will be within ± 5% (far short of the 100% required in the highest band or the 95% required in the next highest band).

Table 2.1: Prevalence Bands, Sample Size Selection Criteria, and Candidate Sample Sizes

Prevalence Band					Criterion	Sample Size
10%	≤	Prevalence			**100%** of margins of error within ± **5%**	**2,230** documents
7%	≤	Prevalence	<	**10%**	**95%** of margins of error within ± **5%**	**3,230** documents
5%	≤	Prevalence	<	**7%**	**95%** of margins of error within ± **6%**	**3,400** documents
3%	≤	Prevalence	<	**5%**	**80%** of margins of error within ± **6%**	**5,080** documents
2%	≤	Prevalence	<	**3%**	**80%** of margins of error within ± **7%**	**7,260** documents
1%	≤	Prevalence	<	**2%**	**70%** of margins of error within ± **8%**	**9,570** documents
		Prevalence	<	**1%**	**50%** of margins of error within ± **10%**	**12,050** documents

With regard to the sample sizes, it may be observed that, as expected, required sample sizes, even allowing for the looser criteria, increase as we get to the lower prevalence bands. Even at the lower prevalence bands, however, sample sizes are not unreasonable or unmanageable, at least for matters involving medium to large document populations. For matters involving small document populations, proportionality considerations may point to the use of smaller sample sizes than those specified. It is reasonable to adopt smaller sample sizes in such cases (while recognizing that the smaller sample sizes may mean accepting a recall estimate with a larger margin of error).

It should also be noted, with regard to the sample sizes, that the sizes specified in Table 2.1 have been derived under the assumption that, at the time of deciding on the size of the Negative Sample, we do not yet know the outcome of the review of the Positive Sample and that our working gauge of overall prevalence is not very precise. If, in any given circumstance, we already know the outcome of the review of the Positive Sample,[70] or if we have more precise information about the prevalence of responsive material in the Review Set at hand, we may be able to arrive at a sample size that is better tailored to the particular circumstances of the matter *sub judice*, and, as a result, may be smaller than the default sizes specified in the table.

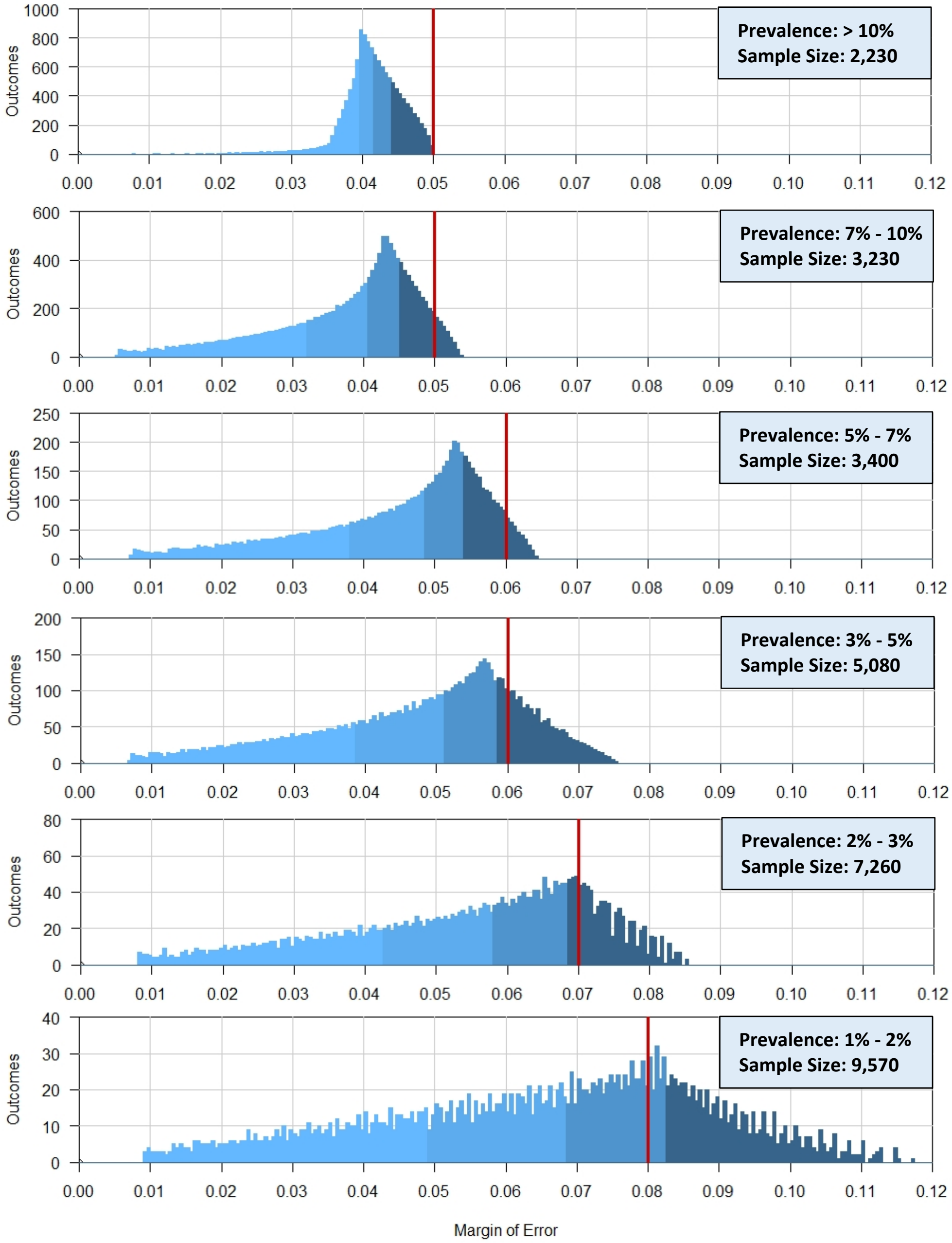
Further analysis of candidate sample sizes

In order to get a more complete view of the power of the candidate sample sizes identified in Table 2.1, we can take a closer look at the distribution of the margins of error each may occasion (if used within its corresponding prevalence band). Figure 2.4 shows, in the form of histograms, these distributions for the sample sizes in the upper six prevalence bands.[71] We show the distribution for the lowest band (prevalence < 1%) in a separate chart (Figure 2.5), as the spread of the distribution for that prevalence band requires a different scale for the horizontal axis of the chart.

[70] It is possible, of course, to sequence the steps of the validation exercise so that this information is available at the time of selecting the size of the Negative Sample.

[71] In the charts in Figures 2.4 and 2.5, the vertical red line shows the margin-of-error threshold operative in the criterion for selection applied for a given prevalence band.

Figure 2.4: Distribution of Margins of Error: Upper Six Prevalence Bands



To supplement the information provided by the histograms in Figures 2.4 and 2.5, we provide, in Table 2.2, the five-number summary (minimum, first quartile, median, third quartile, maximum) for each of the distributions we are considering.

Table 2.2: Five-Number Summaries: Distribution of Margins of Error Occasioned by Candidate Sample Sizes

	Sample	**Distribution of Margins of Error**				
Prevalence Band	**Size**	**Min**	**Q1**	**Med**	**Q3**	**Max**
10% ≤ Prevalence	2,230	± 0.8%	± 3.9%	**± 4.2%**	**± 4.4%**	± 5.0%
7% ≤ Prevalence < **10%**	3,230	± 0.6%	± 3.2%	**± 4.1%**	**± 4.5%**	± 5.4%
5% ≤ Prevalence < **7%**	3,400	± 0.7%	± 3.8%	**± 4.9%**	**± 5.4%**	± 6.4%
3% ≤ Prevalence < **5%**	5,080	± 0.7%	± 3.9%	**± 5.1%**	**± 5.8%**	± 7.5%
2% ≤ Prevalence < **3%**	7,260	± 0.8%	± 4.3%	**± 5.8%**	**± 6.8%**	± 8.5%
1% ≤ Prevalence < **2%**	9,570	± 0.9%	± 4.9%	**± 6.9%**	**± 8.2%**	± 11.8%
Prevalence < **1%**	12,050	± 1.5%	± 7.0%	**± 10.0%**	**± 12.4%**	± 56.1%

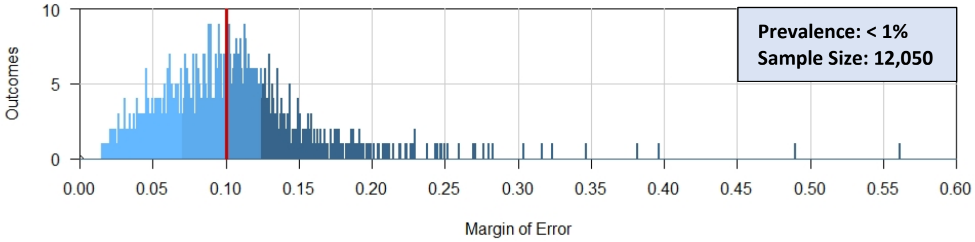
In assessing these distributions, it is important to remember that **a given margin of error is not an end in itself**. Our goal in conducting a validation exercise is to gather evidence that the review has (or has not) succeeded in making a reasonably complete retrieval of the documents responsive to the operative production requests. While our achievement of that goal is furthered by more precise recall estimates, it is also true that the goal can be accomplished even when large margins of error attach to a recall estimate, as long as, allowing for the full range of recall values included by the margin of error, we can still exclude from consideration all levels of recall that would be *prima facie* evidence of an ineffective review process. Put more simply, in assessing the distribution of margins of error made possible by a given sample size, we have to focus on the practical objective, and on what is proportionate in a given circumstance, rather than on a specific number. With that general observation in mind, we can say the following about the distributions enabled by the sample sizes shown in Table 2.2.

With regard to the highest prevalence band (prevalence greater than 10%). For this prevalence band, a Negative Sample of 2,230 will suit the objectives of most validation exercises. As required by the sample size criterion, the margin of error will always be within ± 5%. The median value for the margins of error enabled by this sample size is ± 4.2% and, as can be seen from the shape of the histogram, most values are concentrated fairly close to this median (the first quartile value is ± 3.9% and the third quartile value is ± 4.4%). It should also be noted that if, in any given circumstance, we have reason to believe that the lower limit for prevalence in the Review Set is actually higher than 10%, we may be able to reduce the sample size further. For example, if we have reasonable grounds for believing that the prevalence in the data set under review is at least 15%, we could use a Negative Sample of 1,290 documents and still meet the criterion of 100% of possible margins of error being within ± 5%.

With regard to the middle prevalence bands (prevalence between 1% and 10%). The distributions of the margins of error occasioned by the sample sizes specified for these five bands indicate that the sample sizes, apart from meeting the initial selection criteria, should serve their intended purpose reasonably well. For example, for the sample size specified for the 5% - 7% prevalence band (3,400 documents), the median margin of error is just under ± 5% (to be precise, it is ± 4.9%), the third quartile value is less than a half percentage point above ± 5% (it is ± 5.4%), and the maximum margin of error is ± 6.4%, a value that would still allow a meaningful assessment of the effectiveness of a review process in most cases. As expected, the distributions shift further to the right as we proceed to the lower prevalence bands, but the margins of error occasioned by the sample sizes are not so large as to preclude a party from making a meaningful assessment of the recall achieved by a review process. Even at the very challenging 1% - 2% prevalence band, the specified sample size (9,570 documents) yields, as indicated by the third quartile value, a margin of error within ± 8.2% in 75% of the possible outcomes for this sampling design.[72]

With regard to the lowest prevalence band (prevalence less than 1%). The lowest prevalence band, as can be seen from Figure 2.5, represents a steep challenge for practitioners wishing to gain meaningful information about the recall achieved by a review process while still keeping sample sizes within practical limits. The sample size specified for this band (12,050 documents) does a reasonable job of meeting this challenge but may or may not be proportionate in any particular circumstance. As seen by the median value for the margins of error yielded by this sample size, it satisfies the initial selection criterion of generating a set of margins of error of which 50% are within ± 10%; as seen by the third quartile value, 75% of those margins of error are within a still usable (at least in many cases) ± 12.4%.

Figure 2.4: Distribution of Margins of Error: Lowest Prevalence Band



It should also be noted that if, in any given circumstance, we have reasonable grounds for believing that the actual lower limit for prevalence is, while less than 1%, still not so low as to approach 0%, we may be able to arrive at a sample size that is either smaller or enables more precise estimation (or both). For example, if we have reasonable grounds for believing that the prevalence in the data set under review is between 0.5% and 1%, we could satisfy the criterion applied earlier for the lowest prevalence band (i.e., the requirement that at least 50% of outcomes generate a margin of error within ± 10%) using a Negative Sample of 9,140 documents.

Conclusion

Based on the preceding analysis, we can summarize our recommendations for the size of the Negative Sample as follows.

If a responding party has, in the course of the review, gathered sufficient empirical evidence to obtain a rough gauge of the prevalence of responsive material in the Review Set, it should select a Negative Sample in accordance with the following table.

[72] Limiting, as previously discussed, the set of outcomes to those that result in a recall point estimate of 60% or greater.

Table 2.3: Recommended Sizes for the Negative Sample Used in the Estimation of Recall

Prevalence Band	Recommended Sample Size
10% ≤ Prevalence	**2,230** documents
7% ≤ Prevalence < **10%**	**3,230** documents
5% ≤ Prevalence < **7%**	**3,400** documents
3% ≤ Prevalence < **5%**	**5,080** documents
2% ≤ Prevalence < **3%**	**7,260** documents
1% ≤ Prevalence < **2%**	**9,570** documents
Prevalence < **1%**	**12,050** documents

If a responding party has not gathered sufficient empirical evidence to obtain a rough gauge of the prevalence of responsive material in the Review Set, it should select a Negative Sample of **3,400** documents. This is the size of sample recommended for cases in which we have reasonable grounds for believing that the prevalence is between 5% and 7%, so it represents a reasonable middle ground in the absence of any directional information about prevalence levels. As a practical matter, moreover, drawing and reviewing a Negative Sample of 3,400 documents is not an unduly burdensome task to ask the responding party to perform in the interest of fostering a well-grounded trust in the results of a review. Further, when viewed in terms of its power at constraining sampling error, when no restriction on prevalence is applied, a Negative Sample of 3,400 documents will, as can be seen from the five-number summary of the distribution of margins of error it generates (Table 2.4), be reasonably effective as a basis for meaningfully precise estimates of recall.[73]

Table 2.4: Five-Number Summary: Negative Sample of 3,400 documents, No Control on Prevalence

Prevalence Band	Sample Size	Min	Q1	Med	Q3	Max
0% ≤ Prevalence < **100%**	3,400	± 0.5%	± 3.3%	**± 3.7%**	**± 4.6%**	± 54.3%

A final note. As provided for in the Protocol (and emphasized in the associated Commentary), the sample sizes specified in the Protocol (and given grounding in this chapter) are provided as *default sizes* that may be adjusted as circumstances warrant. It is impossible to specify, in the abstract, sample sizes that will be suitable in every particular circumstance. It is expected that practitioners will take into account the factors that are operative in the case of their specific matter and review process, apply the principle of proportionality,[74] as well as the sound statistical reasoning

[73] Of course, as discussed above, this view of the power of the sample is subject to caveats, given the fact that it mixes high and low prevalence cases and so does not give us a well-focused view of the power of the sample size in any specific circumstance. We are always better served if we can design the sampling exercise with the aid of some, even very rough, estimate of prevalence. Nonetheless, allowing for the absence of a control on prevalence, the summary shown in Table 2.4 does show that the sample size will produce meaningful results in many cases.

[74] I.e., "considering the importance of the issues at stake in the action, the amount in controversy, the parties' relative access to relevant information, the parties' resources, the importance of the discovery in resolving the issues, and whether the burden or expense of the proposed discovery outweighs its likely benefit" (Fed R. Civ. P. 26(b)(1)).

described in this chapter, and then, when those considerations so warrant, adjust the default sample sizes to suit their circumstances.[75]

[75] In making adjustments to sample sizes, practitioners may find it helpful to call on the support of individuals with the appropriate statistical expertise (whether these individuals are in-house experts, support specialists provided by a technology vendor, or independent consultants with a practice in legal discovery).

Chapter 3: A Glossary of Terms of Art Used in Validation

This chapter is intended to provide practitioners with guidance on the meaning and usage of key statistical terms and concepts. It is not intended to be a complete glossary of e-discovery terms.[76] Its focus is limited to terms and concepts related to validation and, specifically, those aspects of a term or concept that are relevant to the implementation of the Protocol's provisions.

Confidence Interval. A confidence interval, like the related concept of a *margin of error*, is a means of quantifying the amount of uncertainty that *sampling error* (i.e., the effect of chance in the random selection process) contributes to a statistical estimate. If a validation exercise tightly controls sampling error (typically through the use of large samples), the confidence interval associated with the estimate will be smaller. If a validation exercise leaves considerable scope for sampling error (typically because it uses small samples), the confidence interval associated with the estimate will be larger.

More specifically, a confidence interval[77] defines a range of values for the population parameter of interest (e.g., for the recall achieved by a review) that would not be incompatible with the results observed in the sample. Values within the range must be entertained as reasonably possible values of the target parameter; values outside the range can reasonably be ignored. For example, if the population parameter of interest is the percentage of responsive documents in a population of 2,000,000 documents, and if, in order to obtain an estimate of that parameter, we draw a sample of 400 documents and find 80 of the sampled documents to be responsive, we can calculate the point estimate for our target parameter to be 20% and the confidence interval associated with that estimate to be the range from 16.1% to 23.9%. Any of the values in that range would not be incompatible with getting the sampling result that we did (80/400) and therefore must be considered reasonably possible as the actual value; any of the values outside of the range, however, would be unlikely to yield the observed sampling result and so can reasonably be dismissed.

To this point, we have characterized the criteria for a given value's being inside or outside of the range demarcated by a confidence interval using imprecise terms of judgment such as "not incompatible with" and "reasonably possible." In constructing an actual confidence interval, we give these inclusion/exclusion criteria precision and objectivity by casting them in terms of statistical probability. More specifically, we construct a confidence interval at a given *confidence level*, meaning that we specify the likelihood that the range demarcated by the confidence interval will contain the true value in terms of a specific probability (e.g., 90% of the time, 95% of the time, 99% of the time). The higher the confidence level, the lower the risk of the interval failing to include the true value (but also the wider the confidence interval); the lower the confidence interval, the higher the risk of the interval failing to include the true value (but also the narrower the confidence interval). In the example above, we constructed the interval (lower limit: 16.1%, upper limit: 23.9%) at a 95% confidence level, meaning the interval was so constructed that it can be expected to contain the true value 95% of the time (or, viewed from a risk perspective, it can be expected to fail to include the true value 5% of the time). Had we chosen to construct the interval at a 90% confidence level, the interval would be narrower (lower limit: 16.7%, upper limit: 23.3%), but we would be assuming a greater risk that the interval did not

[76] For glossaries with that objective, see The Sedona Conference 2020; Grossman & Cormack 2013.

[77] Also see Grossman & Cormack 2013: "Confidence Interval: As part of a Statistical Estimate, a range of values estimated to contain the true value, with a particular Confidence Level" (p. 12).

contain the true value. Had we chosen to construct the interval at a 99% confidence level, the interval would be wider (lower limit: 14.8%, upper limit: 25.2%), but we'd be reducing the risk that the interval did not contain the true value. The specific confidence level used in constructing an interval in any given instance is a matter of user choice, the choice being determined by the individual's information needs and risk tolerance (and, often, by convention as well[78]).

The range demarcated by a confidence interval need not be symmetrical around the point estimate. The segment above the point estimate, for example, may be smaller than the segment below the point estimate. This is one way in which a confidence interval differs from the related concept of a *margin of error*.

The lower and upper limits of the confidence interval associated with an estimate are typically expressed in parentheses following the point estimate, e.g., "20% (16.1%, 23.9%)." The confidence level at which an interval is constructed can be expressed as a qualifier of the term confidence interval: "95% confidence interval" (a confidence interval constructed at a 95% confidence level), "99% confidence interval" (a confidence interval constructed at a 99% confidence level), and so on.

Confidence Level. A confidence level, as a statistical term of art, quantifies the degree of certainty with which a statement about the results of a sample-based test or estimation procedure can be made (e.g., the degree of certainty with which a statement such as *"the recall achieved by the review process is between 74% and 82%"* can be made). It quantifies the degree of certainty as a probability value (e.g., 90%, 95%, 99%), the calculation of which is based on the specific procedures followed in the exercise (sampling design, estimation procedures, etc.) and the specific inputs to those procedures (population sizes, sample sizes, sample observations, etc.). If, for example, we say that the statement *"the recall achieved by the review process is between 74% and 82%"* is made at a 95% confidence level, we are saying that, given the procedures and inputs used to arrive at that range, the statement will be true (i.e., the actual recall achieved by the review will be in the specified range) 95% of the time (or 19 times out of 20); conversely, we are also saying that, given the procedures and inputs used to arrive at the range, we would expect the statement to be untrue (the actual recall achieved by the review will be outside the specified range) 5% of the time (or one time out of 20).

The confidence level is typically set by the user in advance of a validation exercise and the output statement is constructed to meet that specification. The selection of a specific confidence level is based on the user's assessment of the information needs in a given circumstance, the user's tolerance of the risk of error, and the resources available to the user in conducting the validation exercise (all things being equal, realizing a statement with a higher confidence level will require more resources than a statement with a lower confidence level). In validating processes in legal discovery, a confidence level of 95% is typically used, as it contains the risk of error reasonably well while still providing actionably precise information. The Protocol's validation provisions are all designed to enable statements that can be made at this confidence level, but a practitioner who wished could easily adapt the procedures to enable statements at higher or lower confidence levels.

What is perhaps most important for practitioners conducting a validation exercise to remember is that **a confidence level is not a measure of the quality of a review**. A confidence level is simply a measure

[78] The convention, in legal discovery, is typically to use a 95% confidence level when quantifying the sampling error associated with a statistical estimate.

that helps us gauge how much sample-based uncertainty attaches to the conclusions we draw from a validation exercise; in itself, it says nothing about the process being validated. **A high level of confidence can attach to a low estimate of recall just as well as to a high estimate of recall**.

False Negative. When we, whether using a technology-assisted or manual system, classify a set of items using a binary classification scheme (e.g., when we classify a medical patient as positive or negative for the presence of a disease, a suspect as positive or negative for guilt of a crime, a document as positive or negative for responsiveness, and so on), and when we do so exhaustively (so no items are left unclassified), each of the items in the set will fall into one of four possible result categories: (1) classified positive and actually positive, (2) classified positive but actually negative, (3) classified negative but actually positive, and (4) classified negative and actually negative. The term *false negative* refers to outcomes in the third category (classified negative but actually positive).

Put in terms of a document review for responsiveness, the four outcomes are defined as follows.

- **True Positive (TP)**: A responsive document correctly classified as responsive.
- **False Positive (FP)**: A non-responsive document incorrectly classified as responsive.
- **False Negative (FN)**: A responsive document incorrectly classified as non-responsive.
- **True Negative (TN)**: A non-responsive document correctly classified as non-responsive.

We can summarize the results of a classification effort (or the results of a review of a validation sample) with a 2x2 contingency table that, by cross-classifying items on both their classified status and their actual status, tabulates counts of items for each of the four possible outcomes (this is often called a "confusion matrix."). An example is the following.

Table 3.1: A 2x2 Contingency Table Showing Outcomes from a Binary Classification Effort

2x2 Table		**Actual**		**TOTAL**
		Pos	**Neg**	
Classified	**Pos**	n_{11} (TP)	n_{12} (FP)	$\boldsymbol{n_{1\cdot}}$
	Neg	n_{21} (FN)	n_{22} (TN)	$\boldsymbol{n_{2\cdot}}$
	TOTAL	$\boldsymbol{n_{\cdot 1}}$	$\boldsymbol{n_{\cdot 2}}$	$\boldsymbol{n_{\cdot\cdot}}$

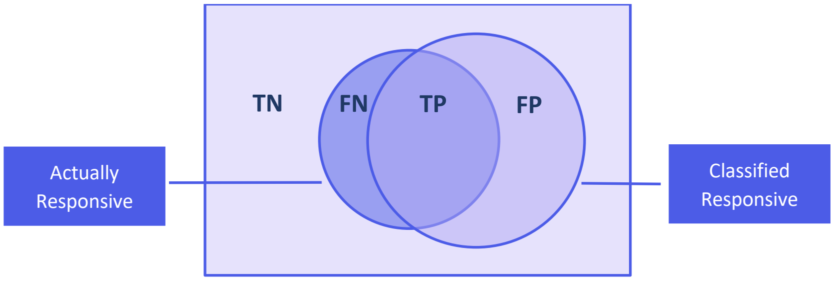
With such a tabulation, calculation of point estimates[79] for the relevant review metrics is straight forward:

- **Recall** = TP / (TP + FN) = n_{11} / $n_{\cdot 1}$;
- **Precision** = TP / (TP + FP) = n_{11} / $n_{1\cdot}$;
- **Prevalence** = (TP + FN) / (TP + FP + FN + TN) = $n_{\cdot 1}$ / $n_{\cdot\cdot}$.

[79] Of course, the point estimates that result from the calculations will be valid only if the counts in the table reflect the results from random sampling of the full population for which the measures are being calculated.

For those who prefer a visual representation, the four-way outcome of a binary classification effort can also be summarized using a Venn diagram, as follows.

Figure 3.1: A Venn Diagram Showing Outcomes from a Binary Classification Effort



As can be seen from the diagram, maximizing recall, where recall = TP / (TP + FN), is a matter of maximizing the True-Positive lens relative to the False-Negative lune; maximizing precision, where precision = TP / (TP + FP), is a matter of maximizing the True-Positive lens relative to the False-Positive lune.

False Positive. See under *False Negative*.

Five-Number Summary. A five-number summary is a convenient and informative way of summarizing the shape and spread of a distribution of values (e.g., the distribution of students' scores on a given test or the distribution of margins of error that may result from a given sample size). Once the values are sorted in ascending order, the summary provides a snapshot of the resulting distribution by reporting the following five data points.

- **Minimum value**: the lowest value in the set.
- **First Quartile value**: the value that marks the cutoff at which 25% of the members of the set are below the cutoff and 75% above the cutoff.
- **Median value**: the value that marks the cutoff at which 50% of the members of the set are below the cutoff and 50% above the cutoff.
- **Third Quartile value**: the value that marks the cutoff at which 75% of the members of the set are below the cutoff and 25% above the cutoff.
- **Maximum value**: the highest value in the set.

In designing a validation exercise and deciding on the sizes of samples, a five-number summary of the distribution of margins of error enabled by a given sample size can provide an insightful view into the power of that sample size.

Histogram. A histogram[80] is a visual representation of the distribution of items in a given population with respect to a variable of interest. In a histogram, the variable of interest is represented on the horizontal axis and "bins" are defined for non-overlapping ranges on that variable (the bins are often defined such that all ranges are of equal size, but this need not be the case). For each bin, a rectangle is then constructed that represents the frequency of items in the range for which the bin is defined (where the frequency may be calculated as simply the number of items in a bin or as the proportion of items in the

[80] For more on histograms, see Freedman et al. 2011: 31*ff*; also: Mendenhall & Beaver 1991: 13*ff*.

bin relative to all other items in the population represented in the diagram). The taller a given rectangle, the more items in the range defined for that bin. For purposes of these guidelines, the histogram offers a convenient way of summarizing the distribution of margins of error that might result from using a sample of a given size.

It should be noted that a histogram is not necessarily the same as a probability distribution. It may be, but only if all items being counted to produce the histogram have an equal probability of occurrence (or of being selected into the sample that is input to the histogram). In the case of the histograms shown in these guidelines, we cannot say that all the outcomes being summarized in a given histogram have an equal probability of occurrence, because there are factors (e.g., the level of recall actually achieved by a review process) that may make some outcomes, within the set of possible outcomes for a given sample size, more likely than others. The histograms therefore should not be read as implying that, for any given sample size, the more possible outcomes that will generate a given margin of error, the more likely the sample size will, in real world practice, result in that margin of error. They should simply be read as showing the full set of possible margins of error that could be generated by a given sample size, as well as the number of outcomes consistent with each margin-of-error span (recognizing that the number of outcomes consistent with a margin of error does not necessarily correlate with the probability of that margin of error).

Margin of Error. A margin of error, like the related concept of a *confidence interval*, is a means of quantifying the amount of uncertainty that *sampling error* (i.e., the effect of chance in the random selection process) contributes to a statistical estimate.

Like a confidence interval, a margin of error defines a range of values for the population parameter of interest (e.g., for the recall achieved by a review) that would not be incompatible with the results observed in the sample. Where a margin of error differs from a confidence interval is that, in the case of a margin of error (unlike a confidence interval), the range of values is always calculated in such a way that it is symmetrical around the point estimate: the distance from the point estimate to the upper limit of the range is the same as the distance from the point estimate to the lower limit. This symmetry allows the range to be conveniently expressed by noting the amount (equal to half the distance between the two end-points of the range) that can be added or subtracted from the point estimate while still remaining within the range of values for the target parameter that are reasonably compatible with the observed sample result. If we take up the example used in the discussion of *confidence interval* (estimating the percentage of responsive documents in a population of 2 million documents, using a sample of 400 documents, and observing 80 responsive documents in the sample) we can express the result using a margin of error as: 20% ± 3.9%. For comparison, the same result is expressed, using a confidence interval, as: 20% (16.1%, 23.9%).

As with a confidence interval, the range demarcated by the margin of error is calculated at a specific *confidence level*, chosen by the user, that gives precise expression (in terms of statistical probabilities) to what is meant by terms such as “reasonably compatible” in any given instance. In reporting a margin of error, it is always important to note the confidence level at which the margin of error was calculated so that those making use of the results can properly assess the amount of sample-based uncertainty that attaches to an estimate.

In the practice of legal discovery, while there are a few circumstances in which making allowance for a range that was not symmetrical around the point estimate would be advisable (and so a confidence interval would be the preferred format for expressing the amount of uncertainty that attaches to an estimate),[81] in the vast majority of cases, the calculation of a range that is symmetrical around the point estimate serves the purpose of expressing sample-based uncertainty perfectly well[82] and so the margin of error, as a simpler (and, for many practitioners, more familiar) mode of expression, is preferred. A margin of error is therefore the default format used in the Protocol and in these guidelines.

Non-Sampling Error. Non-sampling error is a component of the uncertainty that attaches to an estimate. The uncertainty associated with a statistical estimate derives from two sources: *sampling error* and *non-sampling error*. The former results from the operation of chance in the random selection of samples; its impact can be gauged by statistical measures such as a confidence interval and can, to some extent, be controlled by sampling design (e.g., using larger samples). The latter source results from factors external to the random selection process, such as reviewer error (incorrect assessments applied to the items in the sample) or the use of a sample that is actually not random or is not drawn from the full population that is the domain of an estimation exercise.

While discussions of validation exercise tend (rightfully) to give much attention to questions of the measurement and control of sampling error (confidence intervals, sample sizes, and so on), it is important that practitioners also give due attention to the control of non-sampling error, which can sometimes dwarf sampling error. Reviewer error can generally be kept in check by developing a well-articulated set of criteria for responsiveness and by ensuring that the review is conducted under an effective quality control regimen.[83] Checks on the randomness of samples (e.g., through chi-squared tests that compare sample distributions across categories to population distributions across the same categories) can be introduced as a routine step in the sample-selection process. Without adequate control of non-sampling error, the inputs to the methods used to measure and control sampling error may be significantly flawed, rendering those methods substantively irrelevant.

Point Estimate. A point estimate is a single-value estimate (as opposed to the range of values demarcated by a *confidence interval* or *margin of error*) of the population parameter that is of interest in a given exercise. Its calculation is based simply on the observed outcome of the sampling exercise with no allowance for the impact of *sampling error*. Taking up again the example used in the discussion of confidence interval (estimating the percentage of responsive documents in a population of 2 million documents, using a sample of 400 documents, and observing 80 responsive documents in the sample), the point estimate is simply the 20% we obtain from the sample results (80/400). While, given the effect of sampling error, we cannot say that the point estimate is the true value of the population parameter of interest, we can say that it is, as long as the samples that are the basis for the estimate have been properly selected and accurately assessed, an unbiased estimator of the parameter. As such, a point estimate can

[81] Circumstances in which this would be advisable are when the prevalence of responsive documents in the Review Set is extremely low (*e.g.*, less than 0.5%) or when the point estimate for the recall achieved is very high (*e.g.*, greater than 95%).

[82] Apart from cases of extremely low prevalence or exceptionally high recall, an approach that allows for ranges that are asymmetrical around the point estimate will typically result in reducing (relative to the results obtained by a margin-of-error calculation) both the upper and lower limits of the range by around half a percentage point (while keeping the point estimate the same). Practically speaking, differences on that order are not likely to impact the parties' assessment of the effectiveness of a review.

[83] For a discussion of methods to account for reviewer error in the specific context of the TREC Legal Track evaluations, see Webber et al. 2010.

be a useful, and scientifically sound, summary gauge of a population parameter such as the recall achieved by the review process.

The point estimate is most useful when it is considered in conjunction with its associated margin of error (or confidence interval). The point estimate provides us with a useful summary estimate of the value of a validation metric (such as recall). The margin of error provides us with a gauge of the uncertainty associated with that estimate. Together, they enable an informed assessment of the effectiveness of the process under evaluation.

Precision. Precision is a measure of how effective a review effort has been at avoiding overcapture. It answers the question: *Out of all the documents that the review process classified as responsive, what percentage were actually responsive?* The higher the percentage (i.e., the higher the level of precision), the more successful the review has been at avoiding false positives (nonresponsive documents incorrectly classified as responsive); the lower the percentage (i.e., the lower the level of precision), the less successful the review has been at avoiding false positives.[84]

In the validation of a review process, precision is generally of secondary importance to *recall*, as precision (unlike recall) does not offer insight into the completeness of a response to a production request. Even if of lesser priority than recall, however, achieving reasonably high precision serves the interests of both the responding and the producing party. For the responding party, high precision means lower cost for privilege review and production processing as well as reduced risk of producing a document that though non-responsive is damaging in some other way. For the requesting party, high precision means a better-focused production set and thus a quicker path to the documents that really matter.

Prevalence. Prevalence is a measure of the amount of responsive material in a given population of documents (whether that population is the Collected Set, the Review Set, the set drawn from a specific custodian, or a set defined in some other manner). Prevalence is generally expressed as a percentage, with the document as the unit for quantification. For example, a prevalence of 20% in a Review Set means that 20% of the documents in the Review Set are responsive.

While prevalence is not a measure of the effectiveness of a review process, it is (as discussed in Chapter 2 of this document) one of the primary factors determining the sample size required to obtain precise estimates of metrics (such as recall) that are measures of the effectiveness of a review process. It is therefore in the interest of practitioners to obtain at least a reasonable working gauge of prevalence before deciding on the specifics (especially sample sizes) of a validation exercise. Equipped with even a rough estimate of prevalence (which in many cases can be obtained by drawing and reviewing a 600-document simple random sample from the set of documents under review), a practitioner can make better-informed decisions about how to design a validation exercise that, in the practitioner's specific circumstances, is both efficient and meaningful.

[84] For other definitions of precision, see Grossman & Cormack 2013: "The fraction of Documents identified as Relevant by a search or review effort, that are in fact Relevant" (p. 25); see also Sedona 2020: "When describing search results, precision is the number of documents retrieved from a search divided by the total number of documents returned. For example, in a search for documents relevant to a document request, it is the percentage of documents returned from a search that are actually relevant to the request" (p. 354).

Recall. Recall is a measure of the completeness of a review effort. It answers the question: *Out of all the actually responsive documents that reside in the Review Set, what percentage did the review process succeed in identifying?* The lower the percentage (*i.e.*, the lower the level of recall), the less complete the result set; the higher the percentage (*i.e.*, the higher the level of recall), the more complete the result set.[85]

Recall is an essential measure of the completeness of the results of a review process and, as such, is a necessary component of a meaningful validation exercise.

Sampling Error. Sampling error is one source of the uncertainty that attaches to a statistical estimate. It results simply from the fact that the estimate is based on a sample that is a subset of the full population and has been selected by a process that includes chance as a component. The extent to which sampling error can cause departures of a sample-based estimate from the true value of the population parameter being estimated can be gauged by statistical measures such as a confidence interval and can, to some extent, be controlled by sampling design (e.g., using larger samples).

The other source for the uncertainty that attaches to a statistical estimate is *non-sampling error* (i.e., departures of a sample-based estimate from the true value of the population parameter being estimated that result from factors external to the random selection process), the most important of which, in the validation of review processes, is reviewer error. For more on non-sampling error, see the appropriate entry above.

Simple Random Sample. A simple random sample is, in statistical usage, one that has been drawn in such a manner that the specific combination of documents that have been drawn into the sample had the same probability of being selected as did every other possible combination of population documents of the same sample size. Textbook definitions are more precise: "Simple random sampling, or random sampling without replacement, is a sampling design in which *n* distinct units are selected from the *N* units in the population in such a way that every possible combination of *n* units is equally likely to be the sample selected."[86] What this means, in terms of sampling requirements, is that (i) the sample be drawn from the source population in such a way that all documents in that population are available for selection into the sample, (ii) all documents in the source population have an equal probability of being selected into the sample, and (iii) a document, once selected, is not available for re-selection.

This definition of a simple random sample has a few practical implications that should be noted. First, the requirement that every possible combination of the target number of documents have an equal likelihood of being the sample selected means that some methods of selecting the sample, even if intuitively attractive, will not be satisfactory. A method of selection, for example, that specified that every i^{th} document on a list of documents in the source population be selected into the sample would not satisfy the randomness requirement, because, so defined, the selection protocol would *a priori*

[85] For other definitions of recall, see Grossman & Cormack 2013: "The fraction of Relevant Documents that are identified as Relevant by a search or review effort" (p. 27); see also Sedona 2020: "When describing search results, recall is the number of documents retrieved from a search divided by all of the responsive documents in a collection. For example, in a search for documents relevant to a document request, it is the percentage of documents returned from a search compared against all documents that should have been returned and exist in the data set" (p. 360*f.*).

[86] Thompson 2002: 11. See also Freedman et al. 2011 "Simple random sampling means drawing at random without replacement" (p. 340); Mendenhall & Beaver 1991: "Simple random sampling gives every different sample in the population an equal chance of being selected" (p. 120).

exclude from selection many otherwise possible combinations of documents (e.g., combinations that allowed for the inclusion of documents adjacent to each other on the list).

Second, not all document review and analysis tools have the functionalities required for the selection of random samples (although increasingly they do) and not all operators of review and analysis technologies have the expertise needed to distinguish a sample that meets the requirements of randomness from one that does not. A responding party should ensure that it has access to the tools[87] and competencies[88] required to meet the randomness requirement and to test that any samples selected do not depart from the randomness requirement.[89]

Third, given that randomness requires that every possible combination of the target number of documents have an equal likelihood of being the sample selected, any method of screening multiple samples, all of which represent the same state of the result set, to find those that will be favorable for purposes of validation will obviously fail to satisfy randomness. This is because the screening process, by design, excludes samples found to be unfavorable. **Any such screening process would render the results of the validation exercise invalid.**

Stratified Sampling. Stratified sampling is a type of sampling design in which the population is completely divided into non-overlapping segments, or *strata* (sg. *stratum*), and independent samples are drawn from each of those segments. The results of that stratum-specific sampling are then aggregated to obtain aggregate estimates (and associated confidence intervals or margins of error) of the full-population value of the parameter of interest (e.g., recall).

In the validation of review processes used in legal discovery, stratified sampling designs are particularly useful when circumstances are such that the responding party chooses to make multiple productions of responsive documents, meaning that, from a validation perspective, there are multiple Positive Sets (and possibly multiple Negative Sets). Estimation procedures assuming a stratified design may be used even in simple cases (one Positive Set, one Negative Set) when the goal is the estimation of the overall *prevalence* of responsive material in the Review Set. The specific procedures for obtaining estimates (and associated confidence margins of error) when a stratified design is used are discussed in Chapter 1 of this document (in Section 1.4: *Additional Circumstances and Metrics*).

True Negative. See under *False Negative*.

True Positive. See under *False Negative*.

Variance. Variance, in statistical usage, is a measure of the variability to which the members of a given population are subject with respect to a particular feature of interest. When we are using sampling to estimate the value of a given population parameter, we will not be able to know the true population variance (any more than we can know the true value of the parameter of interest), but we can use the sample variance as an unbiased estimator of the population variance. **The estimate of the population**

[87] Among the tools that might be helpful are publicly available random-number generators that can be used to select documents based on Bates numbers.

[88] How to meet the requirement of randomness is a question on which a party could benefit from the support of a consultant with the appropriate expertise.

[89] As an example of a simple test of randomness, one might compare the distribution of documents across custodians in the source population against the same distribution in the sample. There will of course be some differences, but the differences should not be so great that they cannot be attributed to the operation of chance in the sample selection process. A simple chi-squared test can be used to detect departures from what might be attributed to chance.

variance obtained from the sample is not itself the margin of error for a given estimate, but it is an essential element in the calculation of the margin of error (along with the user-specified *confidence level*): the greater the variance, the larger the margin of error.

For a standard textbook treatment of variance, see Mendenhall & Beaver 1991: 30*ff*. For a discussion of variance when estimating a proportion, see Thompson 2002: 39*f*. For a discussion of variance when calculating confidence intervals for recall, see Webber 2013: 9*ff*. For the specific formulae used for the calculation of variance in the validation exercises provided for in the Protocol, see Chapter 1 of this document.

Appendix A: Equation Library

In this appendix we list, in generic form, the equations that are used to obtain the salient metrics, both for validating an exclusionary step (such as applying search terms to narrow the set of documents to be included in the Review Set) and for validating the results of a review process. The fourteen equations included in this "library" represent the full range of operations that are needed to obtain estimates of the proportions and totals of responsive documents in a given set of documents as well as estimates of the summary metrics recall, precision, and prevalence (and the margins of error associated with those estimates). This library thus serves as a reference source for the discussions of specific validation procedures elsewhere in the Guidelines. The library is divided into three subsections: (1) **Notational conventions** (i.e., definitions of the notational devices used in writing the equations), (2) **Building blocks** (i.e., equations for obtaining foundational statistical estimates that are then used to obtain estimates of higher-level metrics), and (3) **Derived metrics** (i.e., equations that, incorporating the output of the building blocks, are used to estimate the summary metrics recall, precision, and prevalence).

Notational conventions

The notational conventions used in the Guidelines are the following.

- N: The number of documents in the population for which a given parameter is being estimated. This is the population from which the sample is drawn.
- n: The number of documents in the sample.
- r: The number of responsive documents observed in the sample.
- p: The estimated[90] proportion of responsive documents in a population.
- t: The estimated total number of responsive documents in a population.
- **Subscripts**. Subscripts are used to indicate the subset of the population for which a given variable is specified; subscripts of note are the following.
    - Subscript $+$: A subscript $+$ indicates that a given variable pertains to the subset of documents assessed as *positive* for either inclusion in the Review Set (in the case of an exclusionary step) or inclusion in the Production Set[91] (in the case of a review for responsiveness). For example, in the case of search terms, the positive (subscript $+$) subset is the subset of documents culled ***in*** by the search terms (i.e., designated for a downstream review for responsiveness); thus N_+ represents the total number of

[90] A caret ("hat") is conventionally used to indicate that a given value represents a statistical estimate (thus $\hat{p}$ represents an estimated proportion, $\hat{t}$ represents an estimated total, and so on. In the interest of readability, we dispense with this use of carets in the equations that follow; it is important to remember, however, that, for most of our validation metrics, we will be arriving at estimates of the true value and not the true value itself. The discussion of the terms in an equation should make clear when a given value is an estimate. (For the most part, anything other than population size, sample size, and number of observed responsive documents (i.e., anything other than N, n, or r) will be an estimate.)

[91] Pending, of course, a review for privilege or other grounds for withholding or redaction.

documents in the Positive Set so defined, n_+ represents the number of documents in the sample drawn from N_+, and so on). In the case of a review process, the positive subset is the subset of documents coded as responsive by the review process (along with associated family members).

- Subscript $\circ$: A subscript $\circ$ indicates that a given variable pertains to the subset of documents assessed as *negative* for either inclusion in the Review Set (in the case of an exclusionary step) or inclusion in the Production Set (in the case of a review for responsiveness). For example, in the case of search terms, the negative (subscript $\circ$) subset is the subset of documents culled ***out*** by the search terms (i.e., designated as not in need of downstream review for responsiveness); thus $N_\circ$ represents the total number of documents in the Negative Set so defined, $n_\circ$ represents the number of documents in the sample drawn from $N_\circ$, and so on). In the case of a review process, the negative subset is the subset of documents coded as non-responsive by the review process (and not associated, by a family relation, with a document coded as responsive).
- Subscript ag: A subscript ag indicates that a given variable represents an *aggregate* value, derived from multiple strata (subsets) of the Review Set.
- Subscript i: A subscript i indicates that the stratum to which a given variable pertains is the i^{th} stratum of the L total strata that contribute to an aggregate value; thus $\sum_{i=1}^{L} t_i$ represents the sum of all the individual stratum total estimates (from stratum 1 through stratum L).

- **Referencing equations.** The equations provided in this library will be referenced via the numbering on the right margin.

Building blocks

- Estimate of a **proportion** (p):[92]

$$p = \frac{r}{n} \tag{1}$$

- Estimated **variance of the proportion** estimator ($var(p)$):[93]

$$var(p) = \left(\frac{N-n}{N}\right)\frac{p(1-p)}{n-1} \tag{2}$$

- Estimate of a **total** (t):[94]

$$t = Np \tag{3}$$

- Estimated **variance of the total estimator** ($var(t)$):[95]

$$var(t) = N^2 var(p) \tag{4}$$

[92] See Thompson 2002: 13, 40.

[93] See Thompson 2002: 40. For a definition of *variance*, see the glossary in Chapter 3 of these guidelines.

[94] See Thompson 2002: 16.

[95] See Thompson 2002: 16

- **Margin of error**[96] associated with an estimate ($M(x)$):[97]

$$M(x) = 1.96\sqrt{var(x)} \tag{5}$$

- **For stratified estimation**: population size ($N_{(ag)}$):[98]

$$N_{(ag)} = \sum_{i=1}^{L} N_i \tag{6}$$

- **For stratified estimation**: estimate of aggregate total ($t_{(ag)}$):[99]

$$t_{(ag)} = \sum_{i=1}^{L} t_i \tag{7}$$

- **For stratified estimation**: estimated variance of the total estimator ($var\left(t_{(ag)}\right)$):[100]

$$var\left(t_{(ag)}\right) = \sum_{i=1}^{L} var(t_i) \tag{8}$$

Derived metrics

- Estimate of **recall**:[101]

$$Recall = \frac{1}{1+\left(\frac{t_\circ}{t_+}\right)} \tag{9}$$

- Estimated **variance of the recall** estimator:[102]

$$var(Recall) = \frac{t_+^2 var(t_\circ) + t_\circ^2 var(t_+)}{(t_+ + t_\circ)^4} \tag{10}$$

- Estimate of **precision**:[103]

$$Precision = \frac{t_+}{N_+} \tag{11}$$

- Estimated **variance of the precision** estimator:[104]

$$var(Precision) = \frac{var(t_+)}{N_+^2} \tag{12}$$

- Estimate of **prevalence**:[105]

$$Prevalence = \frac{t_{(ag)}}{N_{(ag)}} \tag{13}$$

- Estimated **variance of the prevalence** estimator:[106]

[96] Calculated at a 95% confidence level.
[97] See Thompson 2002: 29*f*., 40; see also Mendenhall & Beaver 1991:243*ff*.
[98] See Thompson 2002: 117.
[99] See Thompson 2002: 118.
[100] See Thompson 2002: 119.
[101] See Webber 2013: 5. This formula is an algebraic recasting of the more familiar $t_+/(t_+ + t_\circ)$. The recasting helps with the calculation of the variance (and thus the margin of error) associated with the recall estimate.
[102] See Webber 2013: 14.
[103] For a definition of *precision*, see the glossary in Chapter 3 of these guidelines.
[104] See Feller 1970: 229.
[105] For a definition of *prevalence*, see the glossary in Chapter 3 of these guidelines.
[106] See Feller 1970: 229.

$$var(Prevalence) = \frac{var\left(t_{(ag)}\right)}{N_{(ag)}^{2}} \tag{14}$$

Appendix B: Worked Examples

In this appendix, we walk through two examples of applying the procedures for validating an exclusion made in the basis of search terms and two examples of applying the procedures for validating a review for responsiveness.

Example 1: Validation of search terms (Case 1)

The inputs for our first culling example are as follows.

- The number of documents in the Positive Set (N_+): **300,000**.
- The number of documents in the Positive Sample (n_+): **400**.
- The number of responsive documents observed in the Positive Sample (r_+): **40**.
- The number of documents in the Negative Set ($N_\circ$): **700,000**.
- The number of documents in the Negative Sample ($n_\circ$): **6,000**.
- The number of responsive documents observed in the Negative Sample ($r_\circ$): **25**.

With these inputs, the procedures for validating search terms are executed as follows.

1) Obtain **point estimates** of the target metrics.

    a) Find the point estimate for the number of responsive documents captured by the search terms (t_+).

        i) Using as inputs n_+ and r_+, apply Equation 1 to obtain the estimated proportion of responsive documents in the Positive Set (p_+).

        $$p_+ = \frac{40}{400} = 0.1$$

        ii) Using as inputs p_+ (the output of the preceding step) and N_+, apply Equation 3 to obtain the point estimate for the number of responsive documents in the Positive Set (t_+)

        $$t_+ = 300{,}000 \times 0.1 = 30{,}000$$

    b) Find the point estimate for the number of responsive documents that remain in the Negative Set ($t_\circ$): repeat the steps specified under 1(a), replacing the Positive-Set inputs with the Negative-Set inputs ($N_\circ = 700{,}000$, $n_\circ = 6{,}000$, $r_\circ = 25$).

    $$t_\circ = 700{,}000\left(\frac{25}{6{,}000}\right) = 2{,}917$$

2) Obtain the **margins of error** associated with the point estimates.

    a) Obtain the margin of error associated with the t_+ estimate ($M(t_+)$).

        i) Using as inputs N_+, n_+, and p_+, apply Equation 2 to obtain the estimated variance of the proportion estimator ($var(p_+)$).

        $$var(p_+) = \left(\frac{300{,}000-400}{300{,}000}\right)\frac{0.1(1-0.1)}{400-1} = 0.00022526$$

        ii) Using as inputs $var(p_+)$ (the output of the preceding step) andN_+, apply Equation 4 to obtain the variance of the total estimator ($var(t_+)$).

        $$var(t_+) = 300{,}000^2(0.00022526) = 20{,}273{,}684$$

iii) Using as input $var(t_+)$ (the output of the preceding step), apply Equation 5 to obtain the margin of error associated with the estimate of the total number of responsive documents captured by the search terms ($M(t_+)$).

$$M(t_+) = 1.96\sqrt{20{,}273{,}684} = 8{,}825$$

b) Obtain the margin of error associated with the $t_\circ$ estimate ($M(t_\circ)$): repeat the steps specified under 2(a), replacing the Positive-Set values with the Negative-Set values.

$$M(t_\circ) = 1{,}136$$

3) **Summarize** the result.

a) Responsive documents captured by the search terms: **30,000** ± **8,825**.

b) Responsive documents not captured by the search terms: **2,917** ± **1,136**.

The results obtained in this example would serve as *prima facie* evidence of the effectiveness of the search terms. We estimate that the search terms have captured over ten times (30,000) as many responsive documents as they have missed (2,917).[107] The margins of error are rather large, but that is to be expected at this stage of the collection and review process. As we have emphasized, however, the numbers do not tell the whole story. We must supplement the numbers with a qualitative analysis of the 25 missed responsive documents that the validation exercise has brought to light.[108] If that analysis finds that any important and unique responsive documents are being missed by the search terms, some targeted remediation of the search terms may still be in order.

Example 2: Validation of search terms (Case 2)

The inputs for our second culling example are as follows.

- The number of documents in the Positive Set (N_+): **100,000**.
- The number of documents in the Positive Sample (n_+): **400**.
- The number of responsive documents observed in the Positive Sample (r_+): **20**.
- The number of documents in the Negative Set ($N_\circ$): **1,900,000**.
- The number of documents in the Negative Sample ($n_\circ$): **6,000**.
- The number of responsive documents observed in the Negative Sample ($r_\circ$): **6**.

With these inputs, the procedures for validating search terms are executed as follows.

1) Obtain **point estimates** of the target metrics.

a) Find the point estimate for the number of responsive documents captured by the search terms (t_+).

i) Using as inputs n_+ and r_+, apply Equation 1 to obtain the estimated proportion of responsive documents in the Positive Set (p_+).

[107] Note that, if we were to calculate an estimate of recall (for which we don't advocate, at least in all circumstances), we would obtain a point estimate of 91.1%, with a margin of error (calculated at a 95% confidence level) of ± 3.9%.

[108] Note that the Protocol specifies that the sample size for the Negative Set be large (6,000 documents) not only so that the scope for sampling error is reasonably controlled but also so that a useful number of false negatives, to the extent they exist, are provided for qualitative analysis.

$$p_+ = \frac{20}{400} = 0.05$$

ii) Using as inputs p_+ (the output of the preceding step) and N_+, apply Equation 3 to obtain the point estimate for the number of responsive documents in the Positive Set (t_+)

$$t_+ = 100{,}000 \times 0.05 = 5{,}000$$

b) Find the point estimate for the number of responsive documents that remain in the Negative Set ($t_\circ$): repeat the steps specified under 1(a), replacing the Positive-Set inputs with the Negative-Set inputs ($N_\circ = 1{,}900{,}000$, $n_\circ = 6{,}000$, $r_\circ = 6$).

$$t_\circ = 1{,}900{,}000\left(\frac{6}{6{,}000}\right) = 1{,}900$$

2) Obtain the **margins of error** associated with the point estimates.

a) Obtain the margin of error associated with the t_+ estimate ($M(t_+)$).

i) Using as inputs N_+, n_+, and p_+, apply Equation 2 to obtain the estimated variance of the proportion estimator ($var(p_+)$).

$$var(p_+) = \left(\frac{100{,}000-400}{100{,}000}\right)\frac{0.05(1-0.05)}{400-1} = 0.00011857$$

ii) Using as inputs $var(p_+)$ (the output of the preceding step) and N_+, apply Equation 4 to obtain the variance of the total estimator ($var(t_+)$).

$$var(t_+) = 100{,}000^2(0.00011857) = 1{,}185{,}714$$

iii) Using as input $var(t_+)$ (the output of the preceding step), apply Equation 5 to obtain the margin of error associated with the estimate of the total number of responsive documents captured by the search terms ($M(t_+)$).

$$M(t_+) = 1.96\sqrt{1{,}185{,}714} = 2{,}134$$

b) Obtain the margin of error associated with the $t_\circ$ estimate ($M(t_\circ)$): repeat the steps specified under 2(a), replacing the Positive-Set values with the Negative-Set values.

$$M(t_\circ) = 1{,}517$$

3) **Summarize** the result.

a) Responsive documents captured by the search terms: **5,000** ± **2,134**.

b) Responsive documents not captured by the search terms: **1,900** ± **1,517**.

The results obtained in this example would serve as strong *prima facie* evidence that the set of search terms is in need of expansion (or other forms of remediation). We estimate that, for every 5 responsive documents included by the search terms, about two responsive documents are excluded by the search terms.[109] An included-to-excluded ratio this low means that, in the absence of remediation, we would

[109] Expressed in terms of recall, these results lead us to a point estimate for recall of 72.5%, with a margin of error (calculated at a 95% confidence level) of ± 18.1%. It will be observed that the margin of error associated with the recall estimate is, in this instance, quite large. This is often the case when a broad collection effort has been made, resulting in a very low prevalence of responsive material in the Collected Set. One of the reasons why we do not

have to tolerate a significant loss of responsive material even at the culling stage, when we know, as a practical matter, that a further loss of responsive material will take place at the review stage.[110] There is, moreover, no need to tolerate a ratio this low if we simply allow for lower values of precision.[111] To be sure, the numbers, as always, need to be supplemented by qualitative analysis, but the numbers in themselves are a strong indication that the responding party has more work to do.

Example 3: Validation of a review process (Case 1)

The inputs for our first review example are as follows.

- The number of documents in the Positive Set (N_+): **150,000**.
- The number of documents in the Positive Sample (n_+): **400**.
- The number of responsive documents observed in the Positive Sample (r_+): **320**.
- The number of documents in the Negative Set ($N_\circ$): **1,850,000**.
- The number of documents in the Negative Sample ($n_\circ$): **3,400**.
- The number of responsive documents observed in the Negative Sample ($r_\circ$): **68**.

With these inputs, the procedures for validating the results of the review process are executed as follows.

1) Obtain **point estimates** of the **total** number of responsive documents in both the **Positive Set** and in the **Negative Set**.

a) Find the point estimate for the number of responsive documents in the Positive Set (t_+).

i) Using as inputs n_+ and r_+, apply Equation 1 to obtain the estimated proportion of responsive documents in the Positive Set (p_+).

$$p_+ = \frac{320}{400} = 0.8$$

ii) Using as inputs p_+ and N_+, apply Equation 3 to obtain the point estimate for the number of responsive documents in the Positive Set (t_+)

$$t_+ = 150{,}000 \times 0.8 = 120{,}000$$

b) Find the point estimate for the number of responsive documents in the Negative Set ($t_\circ$): repeat the steps specified under 1(a), replacing the Positive-Set inputs with the corresponding Negative-Set inputs ($N_\circ = 1{,}850{,}000$, $n_\circ = 3{,}400$, $r_\circ = 68$).

$$t_\circ = 1{,}850{,}000\left(\frac{68}{3{,}400}\right) = 37{,}000$$

2) Obtain the **variances** associated with the **total** estimates.

advocate, as a requirement, that responding parties calculate recall estimates (and associated margins of error) at the culling stage is that we do not want to create an incentive for parties to narrow the scope of their collection simply as a way to reduce the margins of error associated with a statistical estimate.

[110] This is true regardless of the review methodology (whether the review is a manual one or some variety of technology-assisted review.

[111] It will be observed that, in this example, the precision of the terms (in the absence of remediation) is already quite low (20/400 = 5%). It is nonetheless also the case that the search terms are "culling in" a relatively small portion of the collected documents (100,000/2,000,000 = 5%), so there is plenty of room for expanding the scope of the terms without losing the value that search-term culling can bring in the first place.

a) Find the variance associated with the t_+ estimate ($var(t_+)$).

i) Using as inputs N_+, n_+, and p_+, apply Equation 2 to obtain the estimated variance of the proportion estimator ($var(p_+)$).

$$var(p_+) = \left(\frac{150{,}000-400}{150{,}000}\right)\frac{0.8(1-0.8)}{400-1} = 0.00039993$$

ii) Using as inputs $var(p_+)$ and N_+, apply Equation 4 to obtain the variance of the total estimator ($var(t_+)$).

$$var(t_+) = 150{,}000^2(0.00039993) = 8{,}998{,}496$$

b) Find the variance associated with $t_\circ$ estimate ($var(t_\circ)$): repeat the steps specified under 2(a), replacing the Positive-Set values with the corresponding Negative-Set values.

$$var(t_\circ) = 1{,}850{,}000^2\left(\frac{1{,}850{,}000-3{,}400}{1{,}850{,}000}\right)\frac{0.02(1-0.02)}{3{,}400-1} = 19{,}699{,}240$$

3) Obtain the **point estimate** for the **recall** achieved by the review process.

a) Using as inputs the total estimates for both the Positive Set and the Negative Set (t_+, $t_\circ$), apply Equation 9 to obtain the point estimate for the recall achieved by the review process ($Recall$).

$$Recall = \frac{1}{1+\left(\frac{37{,}000}{120{,}000}\right)} = 0.764$$

4) Obtain the **margin of error** associated with the **recall** estimate.

a) Using as inputs the total estimates for both the Positive Set and the Negative Set (t_+, $t_\circ$) and their associated variances ($var(t_+)$, $var(t_\circ)$), apply Equation 10 to obtain the estimated variance of the recall estimator ($var(Recall)$).

$$var(Recall) = \frac{120{,}000^2\times 19{,}699{,}240+37{,}000^2\times 8{,}998{,}496}{(120{,}000+37{,}000)^4} = 0.00048716$$

b) Using as input $var(Recall)$, apply Equation 5 to obtain the margin of error associated with the recall estimate ($M(Recall)$).

$$M(Recall) = 1.96\sqrt{0.00048716} = 0.043$$

c) Converting proportions to percentages, summarize the result:

Recall = 76.4% ± 4.3%.

The results obtained in this example would serve as *prima facie* evidence of the effectiveness of the review process. The point estimate for the recall achieved by the review process is greater than 75% and the sample-based uncertainty associated with that estimate has been reasonably constrained (within a margin of error, calculated at a level of 95% statistical confidence, of ± 5%). Of course, as always, it is essential to supplement the quantitative results with qualitative analysis. In this case, that analysis would take the form of an analysis of the 68 false negatives (missed responsive documents) turned up by the validation exercise. Whether the review can be considered complete will depend on whether any of those documents are found to contain information that is both important and unique (not recoverable from documents the review successfully coded as responsive).

Finally, we note that the precision and prevalence estimates for this example are as follows.[112]

- Precision = **80.0%** ± **3.9%**.
- Prevalence = **7.9%** ± **0.5%**.

Example 4: Validation of a review process (Case 2)

For our second review example, we consider a scenario in which the review has been carried out in two phases: one phase in which an initial data set (a large set comprising the more easily collected data) was reviewed and a second phase in which a late-arriving data set (a smaller set comprising the harder-to-collect data) was reviewed. The goal of the validation exercise is to obtain a gauge of the aggregate recall achieved across both data sets; hence this is a circumstance in which we apply procedures suited to a stratified sampling design.[113] The inputs are as follows.[114]

- The number of documents in the Positive Set 1 ($N_{+(1)}$): **150,000**.
- The number of documents in the Positive Sample 1 ($n_{+(1)}$): **400**.
- The number of responsive documents observed in the Positive Sample 1 ($r_{+(1)}$): **320**.

- The number of documents in the Negative Set 1 ($N_{\circ(1)}$): **1,850,000**.
- The number of documents in the Negative Sample 1 ($n_{\circ(1)}$): **3,400**.
- The number of responsive documents observed in the Negative Sample 1 ($r_{\circ(1)}$): **68**.

- The number of documents in the Positive Set 2 ($N_{+(2)}$): **20,000**.
- The number of documents in the Positive Sample 2 ($n_{+(2)}$): **400**.
- The number of responsive documents observed in the Positive Sample 2 ($r_{+(2)}$): **360**.

- The number of documents in the Negative Set 2 ($N_{\circ(2)}$): **480,000**.
- The number of documents in the Negative Sample 2 ($n_{\circ(2)}$): **600**.
- The number of responsive documents observed in the Negative Sample 2 ($r_{\circ(2)}$): **2**.

With these inputs, the procedures for validating the aggregate results of the review process are executed as follows.

1) Following the steps for estimating an aggregate total, obtain the **aggregate point estimates and variances** used in the calculation of recall: $t_{+(ag)}$, $var\big(t_{+(ag)}\big)$, $t_{\circ(ag)}$, and $var\big(t_{\circ(ag)}\big)$.

   a) Obtain $t_{+(ag)}$.

[112] In the interest of space, we do not walk through the steps required to obtain these results. Interested readers are encouraged, however, to do so (following the steps specified in Section 1.4 (*Additional Circumstances and Metrics*), under the heading *Additional metrics*) and to check their results against those given here.

[113] Unless, of course, we chose simply to wait until all the in-scope subsets were reviewed and then conducted the validation exercise (which could then follow the procedures (illustrated in Example 3) for the simple "canonical" case).

[114] It may be observed that, for Example 4, the initial set is the same as the set modeled in Example 3. What Example 4 models in a scenario in which that set is supplemented by a subsequent (separately reviewed) data set of 500,000 documents.

i) Using the stratum-specific inputs $N_{+(i)}$, $n_{+(i)}$, and $r_{+(i)}$, apply Equation 1 and Equation 3 to obtain the point estimate for the number of responsive documents in each stratum ($t_{+(i)}$).

$$p_{+(1)} = \frac{320}{400} = 0.8$$

$$p_{+(2)} = \frac{360}{400} = 0.9$$

$$t_{+(1)} = 150{,}000 \times 0.8 = 120{,}000$$

$$t_{+(2)} = 20{,}000 \times 0.9 = 18{,}000$$

ii) Using as input the stratum-specific total estimates ($t_{+(i)}$ for $i = 1$ to L), apply Equation 7 to obtain the full-population estimate of the number of responsive documents ($t_{+(ag)}$).

$$t_{+(ag)} = \sum_{i=1}^{2} t_{+(i)} = 120{,}000 + 18{,}000 = 138{,}000$$

b) Obtain $var\left(t_{+(ag)}\right)$.

i) Using the stratum-specific inputs $N_{+(i)}$, $n_{+(i)}$, and $p_{+(i)}$, apply Equation 2 and Equation 4 to obtain the estimated variance of the total estimator for each stratum ($var\left(t_{+(i)}\right)$).

$$var\left(p_{+(1)}\right) = \left(\frac{150{,}000-400}{150{,}000}\right)\frac{0.8(1-0.8)}{400-1} = 0.00039993$$

$$var\left(p_{+(2)}\right) = \left(\frac{20{,}000-400}{20{,}000}\right)\frac{0.9(1-0.9)}{400-1} = 0.00022105$$

$$var\left(t_{+(1)}\right) = 150{,}000^2(0.00039993) = 8{,}998{,}496$$

$$var\left(t_{+(2)}\right) = 20{,}000^2(0.00022105) = 88{,}421$$

ii) Using as input the stratum-specific variances ($var\left(t_{+(i)}\right)$ for $i = 1$ to L), apply Equation 8 to obtain the estimated variance of the full-population total estimator ($var\left(t_{+(ag)}\right)$).

$$var\left(t_{+(ag)}\right) = \sum_{i=1}^{2} var\left(t_{+(i)}\right) = 8{,}998{,}496 + 88{,}421 = 9{,}086{,}917$$

c) Obtain $t_{\circ(ag)}$.

$$t_{\circ(ag)} = 38{,}600$$

d) Obtain $var\left(t_{\circ(ag)}\right)$.

$$var\left(t_{\circ(ag)}\right) = 20{,}975{,}506$$

2) Using the output of the preceding step, apply Equation 9, Equation 10, and Equation 5 to obtain the **point estimate** for recall and the **margin of error** associated with that estimate ($Recall_{(ag)}$, $M\left(Recall_{(ag)}\right)$.

$$Recall_{(ag)} = \frac{1}{1+\left(\frac{38{,}600}{138{,}000}\right)} = 0.781$$

$$var\left(Recall_{(ag)}\right) = \frac{138{,}000^2 \times 20{,}975{,}506 + 38{,}600^2 \times 9{,}086{,}917}{(138{,}000+38{,}600)^4} = 0.00042460$$

$$M\left(Recall_{(ag)}\right) = 1.96\sqrt{0.00042460} = 0.040$$

a) Converting proportions to percentages, summarize the result:

Recall = 78.1% ± 4.0%.

The results obtained in this example would again serve as *prima facie* evidence of the effectiveness of the review process. The point estimate for the aggregate recall achieved by the review process across both phases of the review (i.e., across both data sets) is greater than 75% and the sample-based uncertainty associated with that estimate has been reasonably constrained. Again, the quantitative results need to be supplemented by qualitative analysis. In this instance, that analysis would take the form of an analysis of the importance and uniqueness of the 70 false negatives turned up by the validation exercise (68 from Negative Sample 1 and two from Negative Sample 2).

It will be observed that, in this example, the margin of error associated with the aggregate recall estimate was held within ± 5% even though the size of Negative Sample 2 was considerably smaller than that of Negative Sample 1 (600 documents *vs.* 3,400 documents). It was possible to draw a smaller Negative Sample from the supplementary data without having an adverse impact on the margin of error because, in the scenario assumed in the example, the supplementary data set was significantly smaller than the initial data set (500,000 documents *vs.* 2,000,000 documents). This illustrates a general point: **when we are estimating the aggregate recall achieved across multiple data sets, it is not necessary to use the same sample sizes[115] for all data sets.** We can, as appropriate, adjust sample sizes for each component data set in a manner that will still allow us to realize targeted margins of error.

Finally, we note that the precision and prevalence estimates for Example 4 are as follows.[116]

- Precision = **81.2%** ± **3.5%**.
- Prevalence = **7.1%** ± **0.4%**.

[115] Or the default sizes specified in the Protocol. Note that the default sizes specified in the Protocol are selected based on the coverage they provide in the simpler review scenario, that in which there is just one Positive Set and one Negative Set. For more on sample size selection, see Chapter 2 of this document.

[116] Again, in the interest of space, we do not walk through the steps required to obtain these results. Interested readers are, however, encouraged to do so.

References

In re: Biomet M2a Magnum Hip Implant Products Liability Litigation, MDL 2391, Cause No. 3:12-MD-2391 (N.D. Ind., South Bend Division, Apr. 18, 2013).

In re Diisocyanates Antitrust Litig., MDL No. 2862, 2021 WL 4295729 (W.D. Pa. Aug. 23, 2021).

Feller, W. (1970). *An Introduction to Probability Theory and Its Applications*, Volume 1, Third Edition, Revised Printing. John Wiley & Sons, Inc.

Freedman, D., Pisani, R., & Purves, R. (2011). *Statistics*, Fourth Edition. W.W. Norton & Company, Inc. New York.

Grossman, M. R. & Cormack, G. V. (2013). The Grossman-Cormack Glossary of Technology-Assisted Review, 7 *Fed. Cts. L. Rev.* 1.

Grossman, M. R. & Cormack, G. V. (2021). Vetting and Validation of AI-Enabled Tools for Electronic Discovery, in *Litigating Artificial Intelligence*, J. Presser, J. Beatson, & G. Chan, Ed., Edmond Publishing.

Grossman, M. R., Cormack, G. V., & Roegiest, A. (2016). TREC 2016 Total Recall Track Overview. In *TREC*.

Hedin, B., Brassil, D., and Jones, A. (2016). On the Place of Measurement in E-Discovery, in *Perspectives on Predictive Coding and Other Advanced Search Methods for the Legal Practitioner*, J. R. Baron, R. C. Losey, and M. D. Berman, Ed. Chicago: American Bar Association.

Lewis, D. D. (2016). Defining and Estimating Effectiveness in Document Review, in *Perspectives on Predictive Coding and Other Advanced Search Methods for the Legal Practitioner*, J. R. Baron, R. C. Losey, and M. D. Berman, Ed. Chicago: American Bar Association.

Mendenhall, W. and Beaver, R. J. (1991). *Introduction to Probability and Statistics*, Eighth Edition. Boston, MA: PWS-Kent Pub. Co.

Roegiest, A., Cormack, G. V., Clarke, C. L., & Grossman, M. R. (2015). TREC 2015 Total Recall Track Overview. In *TREC*.

The Sedona Conference (2020). The Sedona Conference Glossary: eDiscovery & Digital Information Management, Fifth Edition, 21 *Sedona Conf. J.* 263 (2020).

Thompson, S. K. 2002. *Sampling*, Second Edition. John Wiley & Sons, Inc. New York.

Webber, W. (2013). Approximate recall confidence intervals. *ACM Transactions on Information Systems (TOIS)*, 31(1), 1-33.

Webber, W. and Oard, D. W. (2016). Metrics in Predictive Coding, in *Perspectives on Predictive Coding and Other Advanced Search Methods for the Legal Practitioner*, J. R. Baron, R. C. Losey, and M. D. Berman, Ed. Chicago: American Bar Association.

Webber, W., Oard, D. W., Scholer, F., & Hedin, B. (2010, October). Assessor error in stratified evaluation. In *Proceedings of the 19th ACM international conference on Information and knowledge management* (pp. 539-548).

Acknowledgements

This document has benefited greatly from the many insightful comments provided by David Siffert and his colleagues at NYU's Center on Civil Justice. It has also been greatly improved by suggestions from Samuel Curtis, Nicolas Economou, Niki Iliadis, and Nicolas Miailhe of The Future Society. As observed in the *Notes on Preparation of these Documents*, a Review Group has been very generous with their time in reviewing and commenting on multiple drafts of these documents. The members of the group are as follows.

- Hon. Benes Z. Aldana (Ret.)
  President & CEO
  The National Judicial College
- Jason R. Baron
  Professor of the Practice
  University of Maryland - College Park
- Julia L. Brickell
  Lecturer, Columbia University
  Board Member, The AI Education Project
  Attorney at Law
- Judge Herbert B. Dixon, Jr. (Ret.)
  Superior Court of DC
- Hon. John M. Facciola
  U.S. Magistrate Judge (Ret.)
- Professor Keith R. Fisher
  Distinguished Fellow
  National Judicial College
- Hon. Paul W. Grimm (Ret.)
  David F. Levi Professor of the Practice of Law
  Director of the Bolch Judicial Institute
  Duke Law School
- Maura R. Grossman, J.D., Ph.D.
  Research Professor, University of Waterloo
  Principal, Maura Grossman Law
- Timothy T. Lau, Sc.D., J.D.
  Senior Research Associate
  Federal Judicial Center
  Research Division
- Christine P. Payne
  Shareholder
  Gunster
- Hon. Andrew J. Peck
  United States Magistrate Judge, SDNY (Ret.)
  Senior Counsel, DLA Piper LLP
- Michael Quartararo
  President of ACEDS & Professional Development
  The Association of Certified E-Discovery Specialists

- MK Sapp
  Associate
  Gunster

The perspective and insights provided by this group has been immensely helpful on both general questions and specific points; we are much indebted to the members of the group for their contributions. Contributors participated in their personal capacity and their participation does not necessarily imply an endorsement, by themselves or the organizations with which they are associated, of all particulars of the final product. Finally, we express sincere gratitude for IEEE's support for the project and, in particular, the interest shown by Lloyd Green, Konstantinos Karachalios, and Alpesh Shah.

Of course, responsibility for any errors belongs to the author.

THE
FUTURE
SOCIETY

867 Boylston Street
Boston, MA 02166