Experts Meeting

Epidemiology: Basic Ideas Applied to Museum Collections

A Report from an Experts Meeting
Organized by the Getty Conservation Institute,
June 15–16, 2015

Jim Druzik and Foekje Boersma

The Getty Conservation Institute

Epidemiology: Basic Ideas Applied to Museum Collections

A Report from an Experts Meeting Organized by the Getty Conservation Institute, June 15–16, 2015

THE GETTY CONSERVATION INSTITUTE
LOS ANGELES

The Getty Conservation Institute
1200 Getty Center Drive, Suite 700
Los Angeles, CA 90049-1684
United States
Telephone 310 440-7325
Fax 310 440-7702
E-mail gciweb@getty.edu
www.getty.edu/conservation

Copy editor: Sheila Berg
Designer: Hespenheide Design

The Getty Conservation Institute (GCI) works internationally to advance conservation practice in the visual arts—broadly interpreted to include objects, collections, architecture, and sites. The Institute serves the conservation community through scientific research, education and training, model field projects, and the dissemination of the results of both its own work and the work of others in the field. In all its endeavors, the GCI focuses on the creation and delivery of knowledge that will benefit the professionals and organizations responsible for the conservation of the world's cultural heritage.

ISBN: 978-1-937433-44-4 (online resource)

Cover: Getty Center, Museum South Pavilion, Gallery S113. © J. Paul Getty Trust

Contents

Executive Summary

Foekje Boersma

In June 2015, the GCI's Managing Collection Environments (MCE) Initiative convened a group of researchers, conservation scientists, and conservators to explore whether epidemiological approaches can be used in the investigation of the causal relationships between objects' mechanical damage and their environment. The objectives of the meeting were to identify the methodology and assess the feasibility of an epidemiology study, to discuss its scope, and to identify areas for potential subsequent collaboration.

The meeting established a visual representation of how our research and collection data may be organized. This "Quality of Evidence pyramid" builds from unfiltered background information up through basic study designs to critically appraised articles, validated predictive models including web-based tools, and, finally, systematic reviews including metastudies (see figs. 10, 11). Currently the majority of research and documentation in conservation and conservation science sits within the layers of unfiltered information. In order to make data comparison and exchange more effective, the need to establish protocols and a common language was expressed.

This lack of protocols for data collecting (including a comprehensive list of observables) was seen by meeting participants as a major gap. Data merging and the vetting and analysis of that merged data are important aspects. A critical review of existing literature is also necessary. It is recognized that even though critically appraised articles are published appropriately in the most widely recognized scientific journals, the language becomes a barrier to the field. Education and dissemination should help vernacularize the literature. Several tools are being developed to make research more accessible and applicable for management purposes (Analytica, Haber's risk assessment tool, IPI's e-climate, V&A RHevents).

Suggestions for an epidemiological study were discussed, and it was established that the focus would be the identification of climate-induced mechanical damage to susceptible materials. Since organic and hygroscopic materials are most affected, objects constructed from these materials would be the main focus of the study. The research can include objects in a wide range of environments, from collections in libraries and archives to churches, historic buildings, and museums. Objects without damage would be included in the study, especially those exposed to one or more climatic events that theoretically would have led to damage yet have not exhibited damage.

Loss of value versus degree of negative change, acceptable rates of damage, and life expectancy are recognized as important considerations for management. However, these considerations are outside the scope of this study, which is intended to remove as many variables as possible.

The strength of epidemiology is that it enables a large sample size, allowing the phenomenon of interest to be consistently identifiable among an array of situational variables. In terms of sample size, a correlation between the layers of the pyramid and the sensitivity of the observational technique was identified. The large groups of observed objects that underpin the pyramid correlate to categorical observations. Increasing the sensitivity of the technique used to monitor objects may reduce the sample size, making studies more feasible. Acoustic emission was identified as a highly sensitive technique that is now ready to be applied in the field. Although interpretation of the data is still difficult, there is an emerging group of experts eager to exchange data.

Meeting participants agreed on the importance of working in collaboration with one another and with other colleagues active in this area elsewhere in the world, leading to a much greater collective impact, especially when one considers the sample sizes required for this kind of research. Although specific formal collaborative projects were not identified, the participants expressed the wish to work together on an informal basis, sharing information and assisting with data analysis and interpretation. Although epidemiological approaches may have been applied to individual projects, it is understood that our field has only recently seen the advantages of synthesizing research on a larger scale and that this approach requires more refined thinking. It will be necessary to conduct preliminary, pilot studies to determine feasibility and to develop a methodology.

FIGURE 1.
Meeting at Windmill Hill on the Waddesdon Estate.
Photo: Foekje Boersma, GCI (J. Paul Getty Trust)

Introduction

Foekje Boersma

The June 2015 meeting of the GCI's Managing Collection Environments Initiative was held at the Rothschild Foundation at Windmill Hill on the Waddesdon Estate in Aylesbury.[1] It provided the perfect setting for two days of intensive brainstorming, not only because of its picturesque location in the English countryside, but also because of its inspiring sustainable architecture, which offers a stable internal environment for the Waddesdon Archive's repositories by the clever adaptation of passive (nonmechanical) climate control features, thus avoiding the need for air-conditioning.

The idea of adapting epidemiology to collections was discussed with colleagues from the National Trust (U.K.), English Heritage, Victoria & Albert Museum, Rijksmuseum Amsterdam, Technical University Eindhoven, Doerner Institut, Fraunhofer Institut, Jerzy Haber Institute, Norwegian Institute for Cultural Heritage Research, Canadian Conservation Institute, Institute for the Preservation of Cultural Heritage at Yale University, and the Image Permanence Institute (see Appendix 2, List of Participants). The meeting, moderated by Sarah Staniforth, president of the International Institute for Conservation of Historic and Artistic Works (IIC), set the stage for sharing experiences and for exploring potential ways to collaborate in this research.

The objectives of the meeting were to identify the methodology and feasibility of an epidemiology study, to discuss its scope, and to identify areas for potential subsequent collaboration. In preparation for the meeting, the group of invited experts was sent a discussion paper that explored the terminology and essential concepts of epidemiology. An updated version of this paper is included in this report (Part 1) and is followed by a summary of the meeting's discussions (Part 2). Specific terms used in this report are indicated with an asterisk (*) at first mention and can be found in Appendix 1, Glossary.

PART 1

Epidemiology Applied to Museum Collections

Jim Druzik

Epidemiology: A branch of medical science that deals with the incidence, distribution, and control of disease in a population.

Introduction to Epidemiology

Epidemiology* is the study of the distribution of a disease or a specific adverse condition in a targeted population. Applied to cultural heritage, epidemiological methods can identify how a physical condition or environmentally driven adverse effect is distributed in museum collections and quantify the efficacy of passive responses or active treatments. In the conservation context, epidemiology involves both individual studies of comparable groups of artifacts and a governing set of principles for evaluating the reliability and accuracy of evidence. Thus it may be used to develop rational guidelines for collection environments, with the idea of identifying safe indoor recommendations with respect to temperature and relative humidity fluctuations, and it may also be considered a governing principle for the GCI Managing Collection Environment (MCE) Initiative.[2]

A review of various twentieth-century models in human epidemiology suggests that these study designs can be applied to museum collections as well, as has previously been noted by a number of researchers (Koestler et al. 1994: 149–64; Suenson-Taylor, Sully, and Orton 1999: 184–94). One main difference is that objects cannot report useful data in the same way patients can; the conservator must therefore perform this function for the object.

This discussion paper describes epidemiology as a concept and briefly compares collection surveys, risk assessments, and epidemiological studies. It also describes features unique to the epidemiological approach that give it an advantage in considering environmental conditions for museum collections, using a few existing examples of applying epidemiology to heritage collections. Four study designs used in analytical epidemiology* and how they establish association and infer causation are discussed: randomized control trials*; cohort studies*; case-control studies*; and cross-sectional studies*. Some of the challenges are presented, such as quality of evidence (QOE)*, heuristic rules*, and confounding variables* and interactions. Some ideas for a pilot study are also explored.

FIGURE 2.
Getty Center, Museum South Pavilion, Gallery S113
Photo: J. Paul Getty Trust

Collection Surveys and Risk Assessments

The collection survey is generally the most streamlined assessment tool since its intent is to be applied to hundreds if not thousands of objects, often to answer very specific questions. It is rapid and relatively economical and requires the least amount of specialized training. Joel Taylor and colleagues have written a great deal on this subject, as have others, such as Suzanne Keene. They note that it is basically observational in nature; quantification is simple, often limited to 3- to 5-point scales; and it has mostly nominal, ordinal, or ratio variables*. Since the purpose of assessment surveys is to reach broad conclusions about the state of preservation and possible causes of deterioration, they are often a strategic snapshot carried out prior to a more comprehensive activity. They are therefore highly subjective and sensitive to situational variables not always under control (Taylor and Stevenson 1999: 19–42; Taylor 2013: 95–106). Examples of these types of variables are storage context, degree of fatigue for surveyors, past experience, expectations, motivations, and cultural factors. One significant limitation to surveys has repeatedly been shown to be reliability. Taylor (2014) has shown with an experiment that the same set of objects can produce different assessment scores from the same group of assessors and that existing efforts to increase reliability do not necessarily have an impact. The introduction of assessment guides did not always improve the average degree of reliability. “Reliability” refers specifically to reproducibility, not accuracy. Accuracy conveys other sets of uncertainties. Collection assessments have an analog in epidemiology in the form of descriptive epidemiology*, but they are not analytical in that they do not typically apply rigorous causal criteria.

While epidemiology is a form of risk assessment, it does not replace any other risk assessment approaches currently in use in our field, since it deemphasizes, if not disregards altogether, both collection context and the relative value of different objects. It is not a tool for strategic planning per se, so it cannot replace a collection survey. Epidemiology needs true representativeness in study groups, and it goes beyond finding risky associations by reducing the causal uncertainty for the distribution of damage in collections. It is designed to compare different collections of objects, objects in differing conditions, or objects at different times, and the independent variable *environment* is understood in its broadest sense but often for our purposes in the narrower sense of climate. It tests for confounding variables and accepts “chance.” It focuses on the strength and weight of evidence (collectively termed “quality of evidence”). It also calls attention to internal risks such as assessor bias*, the proper framing of questions, “wicked environments”*, and a few other systemic variables that are described in the section “Assessment Challenges” below.

Epidemiology in the Field of Conservation: Some Examples

Other than laboratory-based random control trials, there have been few epidemiological study designs applied to heritage conservation. In fact, Reedy and Reedy (1988) specifically noted the lack of hypothesis testing in conservation research. Kimberly and Emley (1933) compared a group of thirty-four deteriorated books from the New York Public Library with duplicates supplied by other libraries throughout the United States, in both rural and urban settings. They measured paper acidity and copper number* and observed the level of general deterioration. A total of 314 volumes from twenty-three libraries representing thirty-one titles dating from a broad time span showed convincingly that urban polluted locations were more detrimental to book paper than were unpolluted sites.

A study by Keene (1994: 249–64) is noteworthy because it clearly tries to apply epidemiological techniques to environmental degradation of excavated iron artifacts. A half dozen treatment options had been published, including storage below 18% RH (Turgoose 1985: 13–18). Keene's study involved 588 objects and studied four treatments in detail, including untreated storage at ambient RH and below 18% RH. A strategy was selected that calculated survival probability* from life tables*. These methods were taken from a standard epidemiological text. Then a series of examinations were conducted at three, four, and eight years. These "cross-sectional" comparisons allowed Keene to determine the prevalence* of continued degradation for each treatment. It was clear that humidity control had no benefit, and only two of the tested treatments provided any benefit whatsoever. A couple of other attempts were made to compare survival rates of freeze-dried leather treatment through Coxian regression, inspired by the work of Keene and Orton at the Institute of Archaeology, University College London (Sully and Suenson-Taylor 1996: 177–81; 1999: 224–31).

Melin and Legner (2014) carried out a study in Gotland of church pulpits and the forms of damage they exhibited. Sixteen churches were investigated, and their polychrome pulpits were scored on six types of damage: mold, insect flight holes, cracks in wood, open joints, craquelure in the paint layer, and paint delamination (Melin and Legner 2014: 94–109). Extensive records kept on fuel purchases were used to look for a correlation between damage and the amount of background heating each church had employed. Although the energy conversion calculations and heating system characterizations were impeccable, damage correlation as the coefficient of determination (r^2)* showed no correlation with five of the six indicators. In the one that did suggest an association, craquelure in paint layers, the coefficient of determination was 0.4. One might have taken this evidence as suggesting that the null hypothesis* was supported, that is, that damage was mostly unexplained by background heating history alone or that other confounding variables weakened the apparent correlations. However, the authors concluded that a relationship between background heating and damage was suggested by the data. Clearly, this is an important study, but many questions remain about the strength of the evidence.

Another retrospective study worth mentioning is the STEP project, which looked at different kinds of leather book bindings of the same book in different libraries (the British Library and the National Library of Wales). Both the environment and the tannin type were found to influence the breakdown path (red rot) of these leathers due to two main chemical mechanisms of hydrolytic and oxidative nature (Larsen 1994: 48–55).

These kinds of retrospective cohort studies (comparisons between historic populations) are surely the most difficult study designs for teasing out causal associations because of incomplete or inconsistent records. Demonstrating their causality is even more problematic when the overall strength of evidence is weak, making it incumbent on researchers to explicitly review the risks of experimental bias. So the supporting pillars of a good epidemiological study—validity of the cohort populations, lack of confounding variables, strength of evidence, and clear causality—are themselves viewable only through the smoke and distortions of history. Studies like that of Brunskog (2012: 30–36) with similar but less influential environmental factors run up against even more difficult problems.

Rohdin et al. carried out a prospective study (looking into the future rather than the past) and focused on people's self-reported contentment, or lack of it, with their environment and its ventilation (Rohdin, Dalewski, and Moshfegh 2012: 164–74). In fact, this is traditional human epidemiology applied to energy sustainability concerns in historic buildings and shows that valuable heritage buildings can suffer from sick building syndrome just like modern ones do. The building had no historic collections.

Finally, Ekelund et al. (2014) describe the ongoing Climate4Wood project, a collaborative research project of the Rijksmuseum, the Universities of Technology of Eindhoven and Delft, and the Netherlands Cultural Heritage Agency. Although not identified as epidemiological (it is termed a “population study”), it easily falls into the category of retrospective descriptive epidemiology. As a large-scale systematic study of the condition of wooden panels (in furniture and in paintings) it has already strengthened the association between damage and construction and, importantly, is making a case that most damage in these objects appears to have happened prior to 1900. The survey is combined with computer modeling and the construction and testing of replicas to help identify the conditions under which damage patterns originally developed in historic artifacts. Surrogates are often used in randomized clinical trials in place of people or objects for which exposure to an adverse effect is ethically questionable or impossible. This makes the Ekelund et al. studies a combined descriptive and analytical epidemiological approach.

Criteria for Judging Causality

Epidemiology stresses how strong an association is and whether it rises to the level of convincing causation. In the 1960s Austin Bradford Hill described nine criteria to use when estimating the quality of epidemiological evidence (see Fedak et al. 2015):

— *Temporality.* Does the exposure or triggering event precede the condition?
— *Strength of evidence.* Is the magnitude of the measured effect great or small?
— *Weight of evidence.* Do independent studies show consistency in their measured results?
— *Dose-response* relationship.* Does the response mirror the size of the exposed dose? (Not all responses are linear or continuous, so this criterion may be hard to verify.)
— *Plausibility.* Does the causal relationship make theoretical sense given what we currently know?
— *Alternative explanations.* When selecting a cause, have all the alternative explanations been fully explored? Often the research stops looking when the first good explanation is found. This is a very strong tendency going up against stereotypes.
— *Specificity.* Is the effect of the cause specific to the response?
— *Coherence.* Does all the evidence present a self-consistent story?
— *Analogy.* Is a causal relationship backed up by experimental evidence?

It is assumed that Bradford Hill’s suggestions are applied by researchers during all phases of research. But the quality of evidence testing is so important in epidemiology that it is fairly common to use committees of experts, formal protocols, and other novel approaches when verifying conclusions. This is particularly true when data integration is applied in unique ways and could result in degrading the value and importance of some criteria (Fedak et al. 2015).

About Carrots, Ecological Fallacy, and p-Hacking

Every statement in the following list is true.[3]

1. Nearly all sick people have eaten carrots, demonstrating a cumulative effect on health.
2. An estimated 99.9% of people who die from cancer or heart disease have eaten carrots.
3. 99% of people who died in car crashes ate carrots in the 60 days prior to their accident.
4. 93.1% of juvenile delinquents come from homes where carrots are served routinely.
5. Among people born in 1839 who later ate carrots, there has been a 100% mortality rate.

The main problem with statements such as these is that no comparative group statistics are provided. Comparative conditions are essential to epidemiology. In addition to the examples just above, Gordis mentions the following statement attributed to the prominent second-century Greek physician Galen of Pergamon: “All who drink of this treatment recover in a short time, except those whom it does not help who die. It is obvious therefore, that it fails in incurable cases.” In this case, the problem is that the statement can apply to any liquid regardless of whether or not it provides a benefit. Similarly, “No objects lasts forever” and “All objects last forever” represent the two forms of verifiability. Sometimes a statement can be presented in order to imply or directly suggest that correlation means causation, when in fact there may be no causation at all. Symptoms and signs are often shared by unassociated causes.

The phrase “Correlation is not causation” has become the centerpiece of what is termed the ecological fallacy*, or ecological inference fallacy. An ecological study is an observational investigation in which at least one variable is measured at the group level. Thus, it might be possible to plot length of crack propagation versus per capita days below a certain relative humidity threshold. This might or might not show a positive correlation; crack propagation may or may not be a true effect. These types of errors, derived from poorly considered evidential inference, are the reason that ecological studies are recommended only for initial explorations into causality and building testable hypotheses.[4] Some sources have termed these associations “nonsense correlations” (International Epidemiological Association 2014) .

The final questionable practice discussed here is that of p-hacking*, or more precisely, post hoc p-hacking. Any metric used to monitor performance ceases in some way to be an effective metric. The importance of the p-value* in estimating the validity of a hypothesis for an adverse cause is explored below. A p-value of 0.05 suggests that a given hypothesis has one chance in 20 of being false. Yet there are many ways researchers can manipulate the data to obtain a low p-value. This is one of the risks inherent in data mining.

Main Study Questions in Epidemiology: What, Where, When, Why, and How

There are several basic study designs in epidemiology divided roughly between descriptive and analytical branches. Under descriptive epidemiology can be found case reports, case series, incidence* reports, and cross-sectional, cohort, and case-control designs. They can also focus on individual or whole populations or groups.[5] These are used to generate hypotheses and answer what, where, when, and to which objects or individuals adverse effects are being found associated. Analytical studies test hypotheses and answer why and how effects occur. Cross-sectional studies assemble and estimate the prevalence of characteristics of interest of a collection at one moment in time. There may be no comparison groups because an earlier "snapshot" of the population or its member is required to provide the basis for comparison. An incidence assessment looks at a collection after an event of major interest, like Superstorm Sandy, and case studies are similar to such studies in conservation. Since many descriptive studies have familiar variations in conservation assessments we need not dwell on them. However, a few study designs have unique features and are described below in more detail. They are randomized clinical trials (RCT), cohort studies, and case-control studies.

The first study design is the randomized control trial (RCT), which is sometimes termed a randomized clinical trial. In the study of humans this is considered the gold standard. Trials begin with a representative population of objects, which are then randomly assigned to two conditions expected to show a measurable difference. After an appropriate period of time, the objects are examined for an effect. As figure 3 shows, a greater number or larger percentage of adverse effects would be expected to show up in the group exposed to "problematic conditions." But this may not always be true since the history of the group or individual will modify their future material responses. If the collection has been exposed in the past to fluctuations long enough for all objects to fully respond, the light-colored box shown below ± 20% may be small or nonexistent (Michalski 2014).

Ethical questions arise when exposing otherwise healthy objects to potentially damaging conditions. Many collection caretakers, however, are skeptical of evidence drawn from replicas, no matter how much care is taken in their fabrication. Nevertheless, the randomized control trial is ideal for laboratory-based studies of model structures, examining dose-response relationships, and other Bradford Hill criteria. Much can be done to reduce the

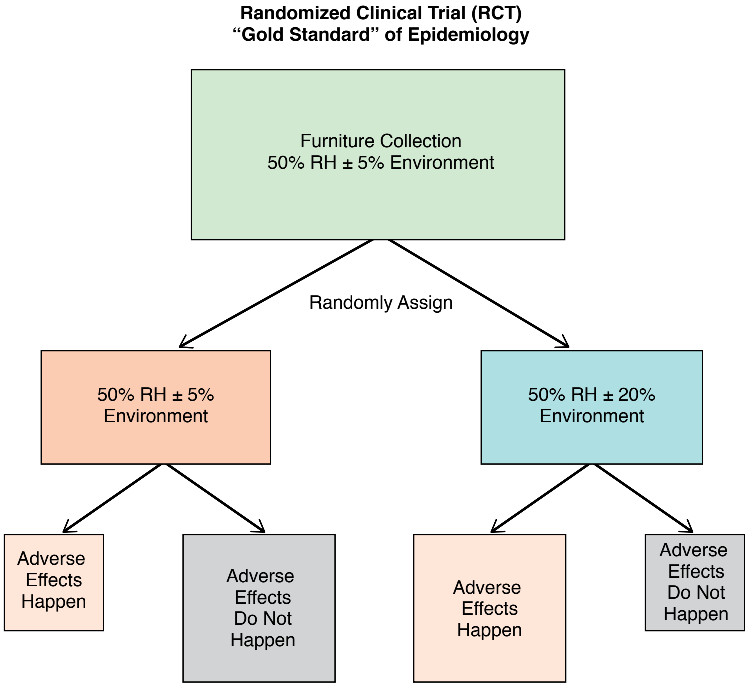


FIGURE 3.
An example of a randomized control trial.

arguments against the reliability of findings based on surrogates, but this often becomes a cultural or sociological challenge more than a materials science one.

Cohort (prospective or retrospective*) study designs are more promising than random control trials in the sense that they do not place objects in harm's way. Because the objects have already come from different environments that meet the experimental design needs, randomized segregation into groups is not necessary. Cohorts may be many and their contents small. There are few limits to the ways cohorts may be defined. The difference between a prospective study and a retrospective study is defined by when it begins and ends. For a retrospective study,[6] the terminus is the present, and we rely principally on historical records to determine the nature of the event or events under investigation. This makes a determination of the reliability, consistency, and accuracy of the historical record important, and gaps need to be filled or at least acknowledged. For a prospective study, the beginning is the present and the terminus is in the future. Documentation can be designed at the onset and protocols established to ensure reliability, consistency, and accuracy. It is also possible to combine historical evidence with future documentation, as is being done in the Climate4Wood project.[7]

Figure 4a presents an example in which two very similar collections of objects with different environmental histories are compared. One might expect that, depending on the adverse effects we are testing, more damage will be found in the less stable environment but that one or both collections have experienced these conditions earlier in their history and remain as they were. The possibility is high in cohort studies that there be no differences to detect. In other words, the same number of total adverse effects happened to the objects on the left as to those on the right. That is exactly what we might expect under the ideal "proofed fluctuation" hypothesis. What we define as new damage would be concentrated in areas of old damage that has been restored. Where once stress concentrations may have been relieved or redistributed, restoration could, if it reestablished those stress patterns, introduce a repeated adverse response.

In the case mentioned above, do we actually need to switch objects and environments? A related kind of study is a modified cross-over study*, shown in figure 4b. It asks the question, "If these objects are stable in their own environments, will they tolerate each other's environment?" The relevance of this question is immediately obvious: it assesses risk concomitant with traveling exhibitions or when applying conservation heating* to modify

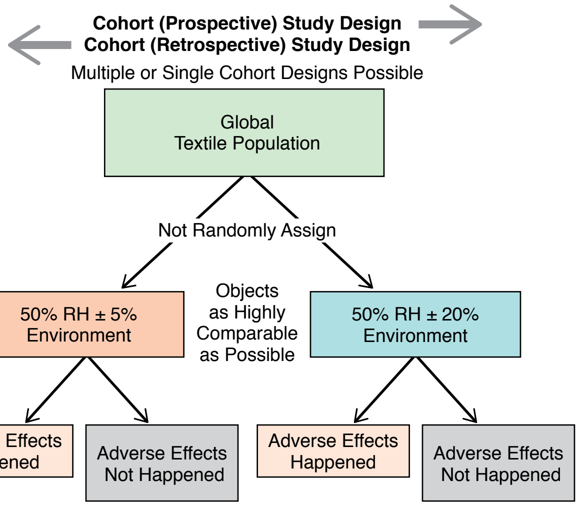


FIGURE 4A. (LEFT)
An example of a cohort study design in which two similar collections of objects with different environmental histories are compared and a hypothesis of what happened is given.

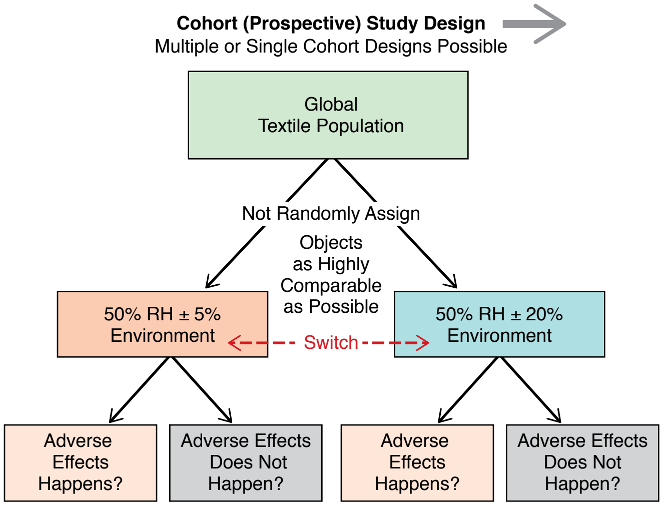


FIGURE 4B. (RIGHT)
An example of a modified cross-over cohort study design, in which the question is raised what would happen if these collections were to switch environments? Observation of the effects could potentially inform a hypothesis. It might be sometimes better to reason from observation rather than test a hypothesis.

an existing environment in order to emulate one that may never have existed before in this location and for these artifacts. For the reasons given above for RCTs, it would be difficult to get a conservator to authorize a cross-over comparison when the results are expected to be new damages, unless the overall benefits are very high. One variation on this theme might be a targeted "incident" survey of condition reports for objects that have been sent on loan to other institutions.

The third study design, the case-control study (fig. 5), begins with objects that possess some defined condition or "disease" (called cases) and objects without a defined condition or "disease" (the "controls"). As in other comparisons, the objects are as well matched as is feasible. But now one works retrospectively to extract which artifacts in both sets were or were not exposed to a defined condition. With the current practices in conservation for recording environmental events, the specificity to detect anomalous exposures on an object-by-object basis, while often discussed, is seldom practicable. Exceptions are when wooden objects may be located near duct openings or exterior walls in harsh climates or seasons, but even then precise measurements may be lacking.

Taking these model designs a step further, we ultimately want to separate out the risks attributable to incorrect temperature and relative humidity from the background risks,

FIGURE 5.
A case control study.

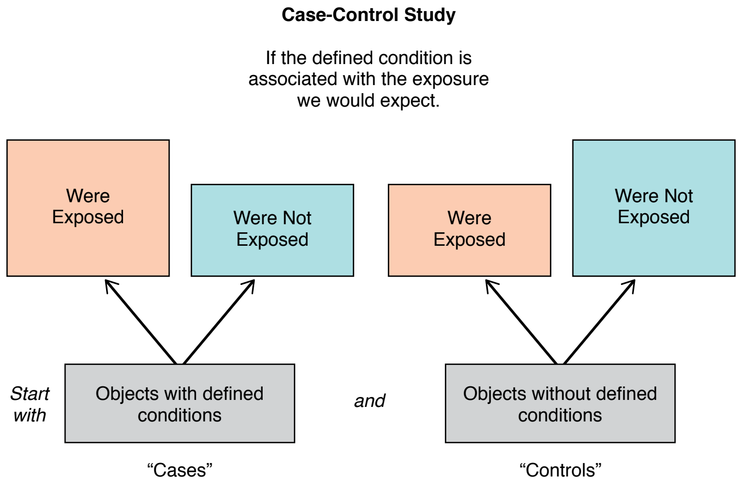
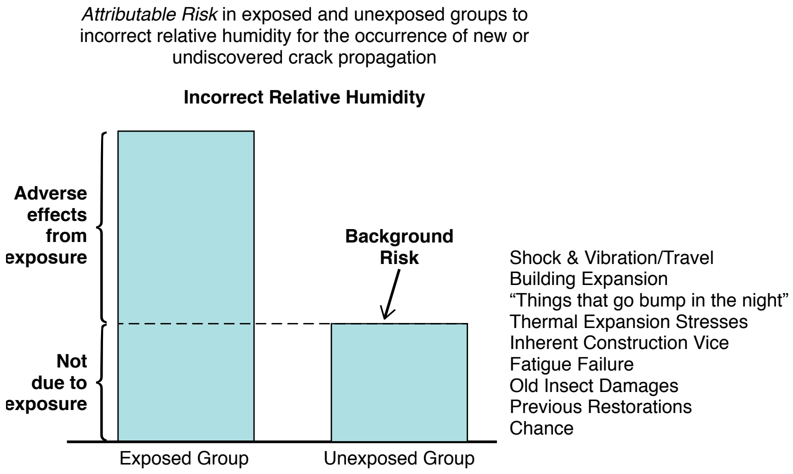

Case-Control Study
If the defined condition is associated with the exposure we would expect.
Were Exposed
Were Not Exposed
Were Exposed
Were Not Exposed
Start with
Objects with defined conditions
and
Objects without defined conditions
"Cases"
"Controls"

FIGURE 6.
Adverse effects from exposure separated from background risks.


Attributable Risk in exposed and unexposed groups to incorrect relative humidity for the occurrence of new or undiscovered crack propagation
Incorrect Relative Humidity
Adverse effects from exposure
Not due to exposure
Background Risk
Shock & Vibration/Travel
Building Expansion
"Things that go bump in the night"
Thermal Expansion Stresses
Inherent Construction Vice
Fatigue Failure
Old Insect Damages
Previous Restorations
Chance
Exposed Group
Unexposed Group

expressed in figure 6. Whether we can do this remains to be seen. However, the main goal is not only to quantify these risks but also to reduce their uncertainty, that is, improve accuracy. That may be the best that one can accomplish. Here we have the same objectives as the risk assessment protocols of Michalski or Waller, although details in all three approaches have significant procedural distinctions that are beyond the scope of this paper.

Assessment Challenges

Virtually every epidemiology textbook contains a chapter on major risks that human bias poses in experimental design and evaluation. The emphasis on risk of bias is much greater in epidemiological literature than in the conservation literature. Kahneman (2011) treats the subject extensively from a decision-making perspective, and Huron (2000) lists sixty versions of bias, many common to clinicians and probably conservators. Only a brief list is included here to convey a sense of how bias permeates intuitive thinking and unconsciously substitutes itself for rational problem analysis.

— *Confirmation bias.* The tendency to interpret new evidence as confirmation of one's existing beliefs or theories.
— *Endowment bias.* Resistance to giving up long-held beliefs.
— *Availability bias.* The favoring of a cause-effect relationship simply because it is more likely to come readily to mind.
— *Law of Small Numbers.* The tendency to generalize a hasty conclusion from small amounts of data. Averages apply to large numbers. A small number of unfavorable incidents does not necessarily represent a systemic trend.
— *Diagnostic momentum.* Once a belief in a cause-effect has been established, regardless of whether it is true, it becomes sticky and hard to reverse.
— *Know-nothing bias.* The tendency to assume that because some issues are unknown, nothing is known.
— *Search satisficing.* The tendency to stop looking for evidence once a reasonable solution is found.
— *Overconfidence bias.* A tendency that is expressed by experts or individuals who feel their experience is sufficient to analyze problems.
— *Representativeness (stereotyping).* The habit of attributing properties to classes of objects and people whether they are known to possess those properties or not.

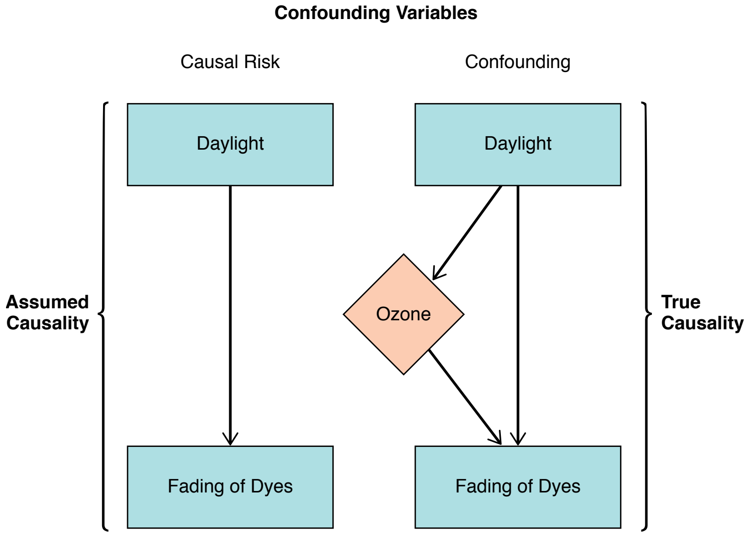
How a question is framed can predispose an individual's attitudes and conclusions toward a given answer. It may even predispose one to either risk acceptance or risk aversion. For example, we can frame an inquiry into the probability of mechanical damage to an object as a result of incorrect temperature and relative humidity. The answer *may not* admit that other factors such as shock and vibration or inherent construction or older restorations can play a role. But if asked to list the common sources of mechanical damage it probably would include other factors, and the more time allowed to consider the question, the longer the list of competing processes tends to get. Either way, each framing may suffer from base rate neglect*. That is to say, one can temporarily ignore or underestimate the frequency of occurrence of other known complicating factors and assume the favored reason is the most probable outcome, leading to overrating its frequency. Yet if the question is framed as an inquiry into the incorrect *distribution of thermal and hygric gradients*, then one is steered away from thinking that an overall simplistic response can explain damage and begins to be more open to other ideas. Being forced to move from a customary way of visualizing a problem can allow better solutions to stand out.

When evaluating causes of environmental damage it might be useful to remind oneself that many causes evoke similar symptoms. This means that intuitive patterns can be misleading. Shanteau (2001) calls these “zero-validity environments”*, while “high-validity environments”* tend to be populated with decisions surrounded by clear and unique cues. The causes themselves are located within these two environments.

Confounding Variables and Interactions

An observed association could turn out to be an incomplete or even false one. It is often assumed that fading is usually a product of direct light–induced damage. But under the proper conditions another agent, ozone, for example, can have a significant or even a larger direct role than light (fig. 7). And there are plenty of reasons why “dark fading” could be confounding these observations as well.

FIGURE 7.
An example of confounding variables.



Epidemiology Pilot Study

At this point it is worth considering a few details of what a heritage-based epidemiological pilot study might look like. This may include the following:

1. A standard cohort study with comparable collections from different sites and environmental conditions.
2. A cross-sectional study fitted to a single existing collection that is soon to undergo a major change in its environment.
3. A small RCT on mock-ups to develop correspondence with published literature.
4. A small set of artifacts in a cross-sectional study to work through a set of common documentation protocols or to test a novel one such as high-frequency photography or acoustic emission monitoring.[8]

5. Linking existing studies.
6. Joining multiple databases to enhance the value of smaller descriptive investigations.[9]

Designs 1–4 are prospective; designs 5 and 6 are essentially retrospective.

Sample considerations

When addressing sample statistics, one may ask, "How many artifacts must be used in our control and exposure groups?" This is not a simple question where one answer may fit most cases and it can be looked up in a table (Alreck and Settle 1995: 470). Sampling decisions in epidemiology often require some advanced estimates of the frequency of a disease and the ability to measure its symptoms. Rational hypotheses are developed and tested against the null hypothesis, which states that no connection exists with the statistical design of the experiment under investigation. The larger the number of hypotheses, the larger the sample sizes will tend to be. Other considerations include economic limits and time restraints.

The biggest challenge present in the above-mentioned studies is related to the quality of recorded evidence with respect to sample statistics and accuracy of observations. Sampling quantities in epidemiology require some reasonable estimate of the results for what is to be measured in advance. Estimation of a sample size for studies based on the comparison of groups of objects exposed to different conditions requires the following five conditions (Gordis 2014: 392):

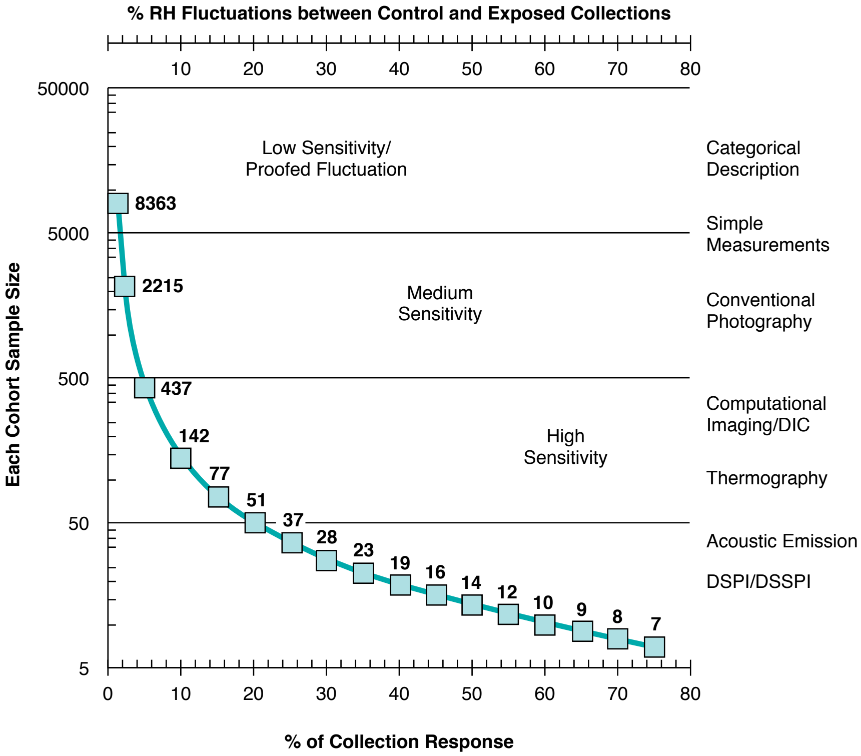
1. The difference in response rates to be detected
2. An estimate of the response rate on one of the groups
3. The level of statistical significance (α, $p<0.05$, Type I error)
4. The value of the power desired (1-β, Type II error)
5. A decision about whether it will be a one-sided or two-sided test

The first condition ensures that one will be able to estimate the difference in the response rates of the groups that are being measured, while the second condition ensures that an estimate of one of the groups is possible. After those two conditions are fulfilled, one should then specify the level of statistical significance. In the third condition, the alpha or p-value is set to 0.05, meaning that there is only a 1 in 20 chance that the two comparison groups do not statistically represent the populations from which they are drawn. This also reduces the probability of making a Type I error, that is, finding that there is no difference between comparative groups but concluding that a difference does exist, thereby rejecting the null hypothesis when it is correct. The fourth condition, 1-β, sets the mark high for avoiding a Type II error, that is, deriving real difference between two groups where we conclude there is no difference. In this case, one fails to reject a false null hypothesis. The fifth condition tests equally for both a positive and a negative effect, or whether a one-sided or two-sided test is going to be performed.[10]

There is no fixed rule for a specific level of statistical significance, but 1 in 20 is common for many disciplines in scientific experiments. It is easy to estimate that at 50% RH ± 2% there will be little or no response in wooden objects for a control cohort, but the question remains, what do we choose for the ± 10% or ± 20% RH cohort? Some help is offered by standard commercial and publicly available computational techniques. The technique used in the discussion below is OpenEpi.[11]

In figure 8 we assumed that one cohort group would serve as the control (50% ± 5% RH); the other was allowed to vary through the largest possible ranges of observable change, plotted as a percent of the group. (We could also envision a cross-sectional study with a single cohort but with specific steps in relative humidity. This would also half the number of objects needed since the first examination would establish the base level of response.) The initial equilibration stage at 50% RH ± 5% would act as the control condition. We also assume that there is adequate time for full response by all objects. The statistical value for α, or the more commonly known p-value, was set to 0.05. The power, 1-β, was set to 0.80, and we tested for two-sidedness. For example, holding the control group unchanged, if only 1% of the exposed collection would show a measurable adverse effect the number of samples in each group would have to be 8,363. In practice, that is an impossible number of objects to analyze. But as the response difference between the two groups grows larger (as the control condition was exposed to more extreme values) the number of required samples per group (to meet our statistical criteria) becomes smaller.

FIGURE 8.
Calculated from OpenEpi: 1-α = 0.95, 1-β = 0.80, two-sided. From Kelsey et al. 1996: tables 12–15.



Our calculations were based on *% of collections response* and not the *% relative humidity fluctuation* but both share an uncommon similarity in their numerical values and we include them together to advance the discussions to follow. The x-axis in figure 6 above is in *% of exposed collection with a measurable change,* but we have added a secondary axis at the top representing the difference in RH fluctuation extremes between the collections to make the important point that sample numbers drop quickly as the % of collection with measurable adverse effects is increased. OpenEpi uses the same methods for calculating the sample size in RCTs and cohort studies with results reported for each of three published standard methods.

There are three obvious ways that we can reduce sample size, as suggested in figure 8. We can make the exposure fluctuations more extreme to induce a larger percent response. But we can also increase the sensitivity of the measurement technique and/or decide to populate our cohorts with a skewed distribution of artifacts thought to represent the most sensitive categories. Some techniques are listed and their sensitivity is ranked on the right edge of figure 8. The positions of the seven groups of techniques do not relate to the graph in any exact manner, nor the positions where the three levels of sensitivity are indicated intended to be taken literally. These are approximations. But it is reasonable to assert that if the assessment technology increases in sensitivity, the result is equivalent to increasing the percentage of artifacts displaying a change, assuming that the detectable change is not "noise" from other causes. Increasing method sensitivity is equivalent to increasing *detectable* collection response, not actual collection response.

Visual observation and ranking is least sensitive, so it would only detect a very large effect. Simple handheld measurements are a little more precise, followed by conventional photography, computational imaging, and then the most powerful tools, acoustic emission, thermography, and laser speckle interferometry. There are dozens of variations on these approaches, but these were selected because they are either commonly used now or have had recent successes in quantifying small changes (Strojecki et al. 2014: 225–32; Krzemien et al. 2015: 544–50)

Combining techniques can increase detection limits to the degree that they each provide unique information. Also, the term "simple measurement" distinguishes between handheld methods and those that occupy a domain beyond sensory detection. This does not mean that handheld measurements are fundamentally inferior. A medium-sized object with dimensions in the 300–500 mm range measured carefully with calipers of appropriate span, accompanied by carefully replicated measurements, can approach 0.05%.[12] Even assuming that measurements of two different individuals have twice the variance, this is still well within the range for a tangential or even a radial cut of wood shrinking and expanding from 40% to 60% RH or vice versa, conditions found in most museums (Hoadley 2000: 110–31).

Detecting a change in artifacts stabilized to long proofing periods is similar to estimating the light sensitivity of colorants. They might once have had an ISO Blue Wool ranking equivalent to BW1 or BW2 but with decades of light exposure have slowed down to barely detectable rates. This serves to enhance our expectation that more sensitive measurement techniques will reduce sample size, but it may not always work this way if object variability is high. In addition to proofing, there is the matter of the type of damage. Is a damage event driven, or does it accumulate in seemingly imperceptible stages? Is it better to consider percent of collection changed as a function of fluctuation or to consider that all objects change continuously from small numbers or cycles to some inevitably large number of fluctuation cycles, after which they can be defined as destroyed?

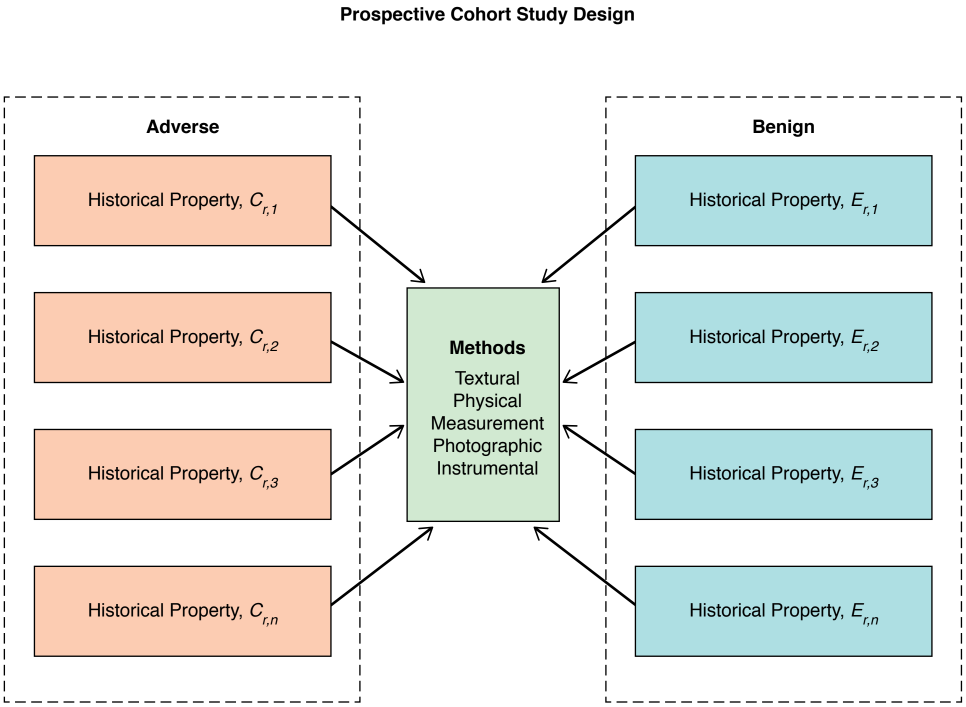
Clearly, at least one pilot study on a limited number of objects should be conducted to refine the right-hand margin in figure 8.[13] Acoustic emission has become recognized as a gold standard in detecting small and microscopic dislocation in wooden objects. These are almost always below visual detection. Of value now is how these signals correlate to physical measurement and the traditional conservator's condition report. The language we use defines the world we see, and the terminology used to describe condition is not immune to the foibles of language, experience, opinion, and overconfidence. This type of information is collectively termed "Categorical Observation" in figure 8. Descriptive reliability has been discussed by Taylor and colleagues, as mentioned earlier (Taylor and Stevenson 1999: 19–42; Taylor 2013: 95–106). Some of the terms used to describe condition are ambiguous. Others infer a primary causality when they could be indicative of other less common causes. Still other terms infer causality but are rationally determined from the visual evidence and may even be described as a weak form of hypothesis derived during examination. Yet few of us have thought carefully about unbiased condition reporting. Rather, we trust in the experience of the conservator making the report. This has traditionally served us well on a case-by-case basis, but it is far from ideal when handling hundreds or thousands of objects or attempting to see small effects reliably. A pilot project should address the conservator's reporting methodologies very carefully. One excellent example offered for painting conservation is that of Appelbaum (2010).

Cohort considerations

In figure 9 it is suggested that the Kimberly and Emley design can be used on artifacts thought to be the most sensitive using delicate techniques but distributed as virtual cohorts. Theoretically, as in the book paper study, the disadvantages of adverse environmental conditions should emerge. This weakness lies in the consistency with which the environmental monitoring is maintained, but the advantage lies in the possibility of assembling cohorts of objects with members that share the closest matches.

The question can also be asked how best to mine the latent information in surviving artifacts that are *the least sensitive*? It was observed in the Getty Foundation's Panel Paintings Initiative that panels that had been sawn through the center and that had cradles applied usually suffered more significantly over time than much thicker panels. Can risks be identified more easily when also examining objects with as yet unidentified group patterns? By adding a prospective multiple cohort study to a set of highly targeted retrospective cross-sectional, incident, or population studies as described above, it may be possible to significantly strengthen the body of evidence already derived from laboratory studies on wood, gesso, and paint on canvas.

FIGURE 9.
An example of a prospective cohort study design.


Prospective Cohort Study Design
Adverse
Historical Property, $C_{r,1}$
Historical Property, $C_{r,2}$
Historical Property, $C_{r,3}$
Historical Property, $C_{r,n}$
Methods
Textural
Physical
Measurement
Photographic
Instrumental
Benign
Historical Property, $E_{r,1}$
Historical Property, $E_{r,2}$
Historical Property, $E_{r,3}$
Historical Property, $E_{r,n}$

Fitting the Quality of Evidence Model to Museum Epidemiology

The common models encountered when ranking the quality of evidence in epidemiology have favored some version of the pyramidal structure shown in figure 10. It builds from unfiltered background information up through the basic study designs—descriptive, case-control, cohort, and RCT studies—into critically appraised articles, validated predictive models including web-based tools, and finally systematic reviews including metastudies.[14] These all build on the evidence to demonstrate, and often prove, causality. Note that background information and expert opinion, to which we might add most anecdotal stories, lie at the lowest level of evidential strength.

Figure 11 inserts elements relevant to sustainable environments research being conducted at the Canadian Conservation Institute, the Haber Institute in Krakow, the Victoria & Albert Museum, the Rijksmuseum, the Technical University at Eindhoven, and the Getty Conservation Institute. The conservation field may apply this same QOE approach to all the projects that we are currently operating or that are theoretically in design. Assuming that a more rational set of "best practices" for sustainable museum environmental conditions rests on the strongest evidentiary foundation that we can provide, figure 11 shows how a network of research builds to that objective and more specifically where lacunae in the structure may be found and filled.

Apparent lacunae notwithstanding, Sarah Staniforth wrote in 2014, "There is a perception that the research has not been done to back up our understanding of the response of materials to changing environmental conditions. My experience of reviewing the literature for *Historical Perspectives on Preventive Conservation* for the Getty Conservation Institute's 'Readings in Conservation Series' is that there is a wealth of research yet it takes time and some detective work to locate and read all the papers and grey literature that records the research" (213–17). One way we could address Staniforth's observations is to populate and call attention to structures like these, holding all the papers and gray literature that records our research. The extent to which some levels are more sparsely populated would then be obvious. This could be done without exclusion, and qualitative ranking within layers would then be the domain of the reader (fig. 12).

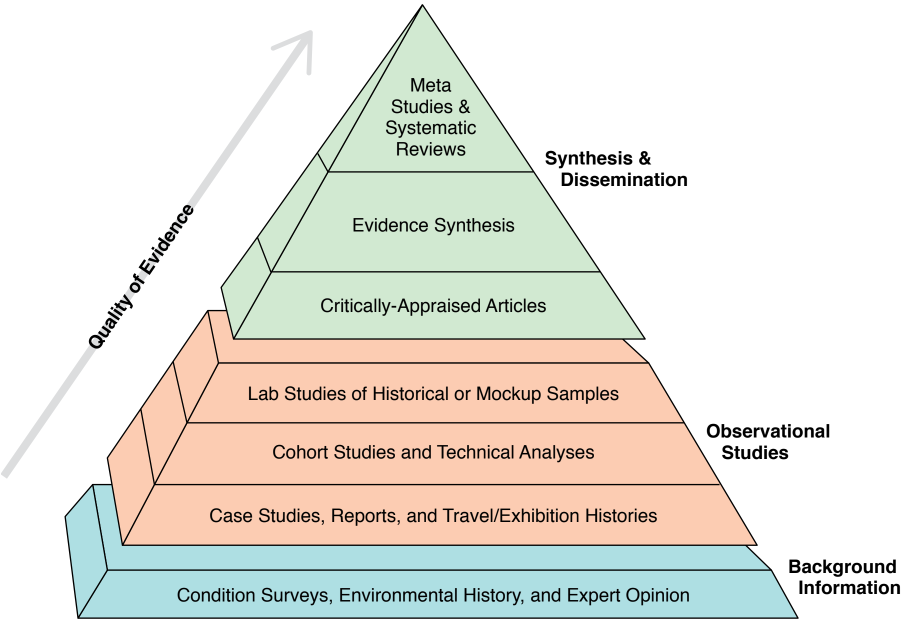

Quality of Evidence
Meta Studies & Systematic Reviews
Evidence Synthesis
Critically-Appraised Articles
Synthesis & Dissemination
Lab Studies of Historical or Mockup Samples
Cohort Studies and Technical Analyses
Case Studies, Reports, and Travel/Exhibition Histories
Observational Studies
Condition Surveys, Environmental History, and Expert Opinion
Background Information

FIGURE 10.
Pyramidal structure ranking the quality of evidence within an epidemiological study in cultural heritage research. Adapted from Kim Hugel, http://www.capho.org/blog/journey-research-levels-evidence (2013).

FIGURE 11.
Adjusted Quality of Evidence pyramid.

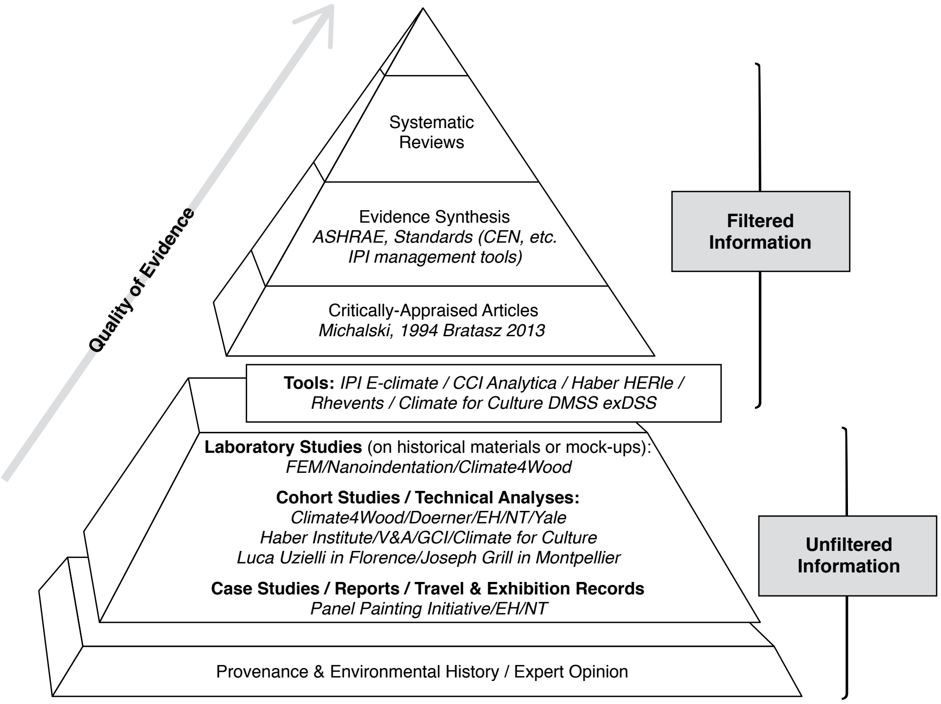
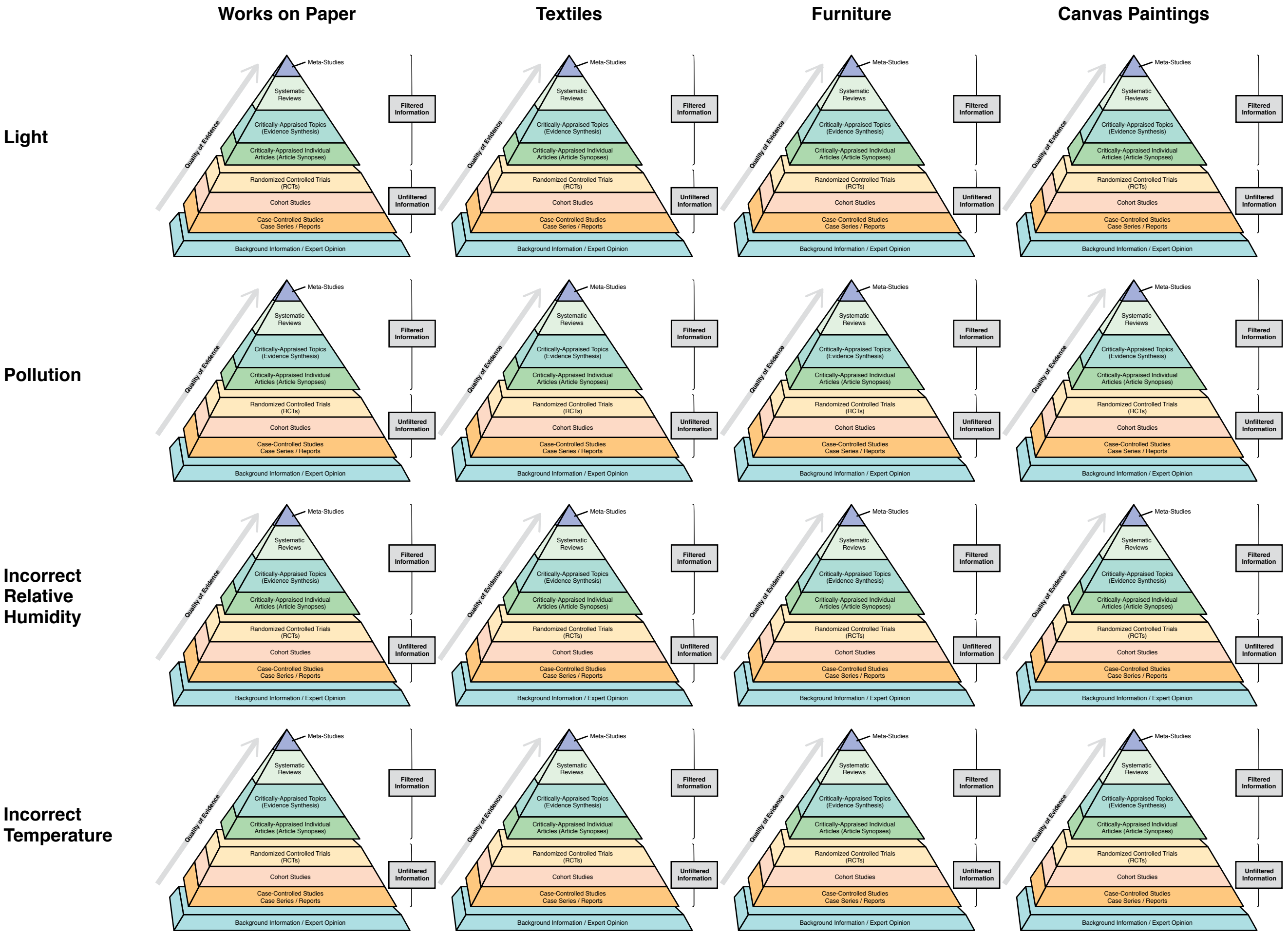

Quality of Evidence
Systematic Reviews
Evidence Synthesis
ASHRAE, Standards (CEN, etc.
IPI management tools)
Critically-Appraised Articles
Michalski, 1994 Bratasz 2013
Tools: *IPI E-climate / CCI Analytica / Haber HERle / Rhevents / Climate for Culture DMSS exDSS*
Laboratory Studies (on historical materials or mock-ups):
FEM/Nanoindentation/Climate4Wood
Cohort Studies / Technical Analyses:
Climate4Wood/Doerner/EH/NT/Yale
Haber Institute/V&A/GCI/Climate for Culture
Luca Uzielli in Florence/Joseph Grill in Montpellier
Case Studies / Reports / Travel & Exhibition Records
Panel Painting Initiative/EH/NT
Provenance & Environmental History / Expert Opinion
Filtered Information
Unfiltered Information

FIGURE 12.
Quality of Evidence revisited:
A proposed Information Matrix.


Works on Paper
Textiles
Furniture
Canvas Paintings
Light
Pollution
Incorrect Relative Humidity
Incorrect Temperature
Quality of Evidence
Meta-Studies
Systematic Reviews
Critically-Appraised Topics (Evidence Synthesis)
Critically-Appraised Individual Articles (Article Synopses)
Randomized Controlled Trials (RCTs)
Cohort Studies
Case-Controlled Studies Case Series / Reports
Background Information / Expert Opinion
Filtered Information
Unfiltered Information

Hypotheses and Evidence

Hypothesis testing* is as important as judging the quality of evidence and lies at the heart of analytical epidemiology. This is especially true in heritage conservation because the questions what, where, when, why, and how artifacts are adversely affected by the temperature and relative humidity of their environment under constantly fluctuating cycles, must be pieced together from the efforts of many research teams. The jigsaw puzzle is abstract and confusing.

The difficulty of “proving” ideas like “proofed fluctuation” has already been addressed. The more likely it is that a historical house museum has experienced a fixed pattern of climate control for several decades, the smaller will be the response to extremes of fluctuations—that is, the effects of black swans* will be dampened—and hence the more difficult it will be to prove proofing.[15] Recall that a 1% or 2% difference means one would have to compare thousands of well-matched case studies—a very difficult task.

But if our ultimate goal is to raise the *quality of the evidence* in support of the concept of proofing, we must reduce the uncertainties surrounding it; help museum professionals to be empowered and more self-confident in its use; demonstrate to facilities managers that small improvements such as increasing the buffering capacity of various microenvironments in buildings are important; and generally give institutions with hard-to-control environments a place at the table in the environmental discussion.

Two important hypotheses to be tested against their null hypotheses follow.

HYPOTHESIS 1. *Improving the QOE can support the proofed fluctuation concept: a suite of three flavors*

— The first proofed fluctuation rule
  - “Proofed value is the largest fluctuation to which the artifact has responded in the past, so any single cycle damage has occurred for this value” (CEN 2010).

— The proofed fluctuation pattern rule
  - “If the future pattern of fluctuation (and pattern of usage) of a permanent collection remains the same as the historical pattern of 30 (or 100) years then the next 30 (or 100) years will cause small to negligible risk of fracture to this permanent collection.”[16]

— Proofed single fluctuation rule (fatigue)
  - “An object with a single proofed fluctuation event is proofed to 100 repetitions at 0.8 of this proofed fluctuation, and 10,000 repetitions at 0.5, before it enters any danger zone.”[17]

HYPOTHESIS 2. *The QOE for Error #4 can be improved*

— “Ignoring the Fact That the Value at Risk Does Not Reside in Original Old Material”: “Old things have certainly seen at least ± 20% RH fluctuations in their time, and often more. If old furniture and paintings appear flawless today, it is only because of restorations. Any original seams or layers vulnerable to ± 20% RH popped long ago. If an old object is actually vulnerable to fluctuations smaller than ± 20% RH it is primarily the re-glued seams and painted in-fills that fell apart, not original material” (Michalski 2016: 27).

This is but a start. Other hypotheses are left to us to develop.

Notes

1 http://www.waddesdon.org.uk/collection/archive-landing.

2 http://www.getty.edu/conservation/our_projects/education/managing/.

3 These examples are taken directly from a lecture presented before judges by Leon Gordis. See https://www.youtube.com/watch?v=3aOEp7MeSV8.

4 http://www.tylervigen.com/spurious-correlations.

5 When whole groups are studied they may be termed "ecological" designs.

6 It is common to hear conservators talk about "anecdotal evidence." This is always synonymous with evidence that occurred in the past, or retrospective.

7 http://www.narcis.nl/research/RecordID/OND1347513.

8 A novel approach was suggested by Matija Strlic in 2014. This incorporated "crowd-sourcing" used in conjunction with image analysis as a means to achieve greater precision and lower uncertainty by creating a semicontinuous visual record of test objects.

9 This is suggested mindful of the complexity presented by metadata and thesaurus compatibilities.

10 Gordis (2014) provides considerably more detail when applying these statistical tools, including useful 2x2 tables for diagramming outcomes from randomized trials with respect to Type I and II errors.

11 OpenEpi, open source epidemiological statistics for public health, available through Emory University, Rollins School of Public Health, http://www.openepi.com/Menu/OE_Menu.htm.

12 This is being tested as part of a study monitoring climate-induced change in a study collection at the GCI.

13 See note 12 above.

14 Metastudies are not studies of metadata in the sense database users may be familiar with. They are studies that use already published data existing in a form that can be combined with other results. Once done, the statistics of combined data sets may be more robust than individual studies alone and answer questions no others could have. For an example of a metastudy on light reciprocity, see Martin, Chin, and Nguyen 2003: 292–311).

15 Significant other properties are that the effects are disproportionate, noncomputable, and psychologically biasing. For our purposes, this may be a hurricane, a major HVAC breakdown, or even a traveling incident.

16 Quote from an unpublished draft paper by Stefan Michalski (2013), "The Power of History in the Analysis of Collection Risks from Climate Fluctuations and Light." Presented at ICOM-CC in Melbourne (Michalski 2014).

17 See note 16 above.

References

Alreck, Pamela L., and Robert B. Settle. 1995. T*he Survey Research Handbook: Guidelines and Strategies for Conducting a Survey.* New York: McGraw-Hill.

Appelbaum, Barbara. 2010. *Conservation Treatment Methodology.* Lexington, KY: CreateSpace.

Brunskog, Maria. 2012. "Paint Failure as Potential Indicator of Cool Indoor Temperature." In *Postprints from the Conference Energy Efficiency in Historic Buildings, Visby, February 9–11, 2011,* edited by Tor Broström and Lisa Nilsen, 30–36. Visby, Sweden: Gotland University Press.

CEN. 2010. *EN 15757: 2010 Conservation of Cultural Property: Specifications for Temperature and Relative Humidity to Limit Climate-Induced Mechanical Damage in Organic Hygroscopic Materials.* Brussels: European Committee for Standardization.

Ekelund, Stina, Paul Van Duin, Andre Jorissen, Bart Ankersmit, Roger M. Groves, Henk Schellen, and Akke Suiker. 2014. "Climate4Wood Museum Study: The Systematic Analysis of Climate-Related Damage on Decorated Wooden Panels in the Rijksmuseum." Paper presented at the

ICOM-CC 17th Triennial Conference, “Building Strong Culture through Conservation,” 15–19 September 2014, Melbourne, Australia.

Fedak, Kristen M., Autumn Bernal, Zachary A. Capshaw, and Sherilyn Gross. 2015. “Applying the Bradford Hill Criteria in the 21st Century: How Data Integration Has Changed Causal Inference in Molecular Epidemiology.” *Emerging Themes in Epidemiology* 12: 14. https://ete-online.biomedcentral.com/articles/10.1186/s12982-015-0037-4.

Gordis, Leon. 2014. *Epidemiology.* Philadelphia, PA: Elsevier Saunders.

Hoadley, R. Bruce. 2000. *Understanding Wood: A Craftsman’s Guide to Wood Technology.* Newtown, CT: Taunton Press.

Huron, David. 2000. “Sixty Methodological Potholes.” http://csml.som.ohio-state.edu/Music829C/methodological.potholes.html. Accessed April 18, 2017.

International Epidemiological Association. 2014. *A Dictionary of Epidemiology.* Edited by Miquel Porta, Sander Greenland, Miguel Hernan, Isabel dos Santos Silva, John M. Last and Andrea Buron. 6th Edition ed. Oxford, United Kingdom: Oxford University Press.

IIC International Institute for Conservation of Historic and Artistic Works and ICOM-CC International Council of Museums–Committee for Conservation. 2014. *Environmental Guidelines—IIC and ICOM-CC Declaration.* https://www.iiconservation.org/sites/default/files/news/attachments/5681-2014_declaration_on_environmental_guidelines.pdf; http://www.icom-cc.org/332/-icom-cc-documents/declaration-on-environmental-guidelines/. Accessed April 18, 2017.

Kahneman, D. 2011. *Thinking, Fast and Slow.* New York: Farrar, Straus and Giroux.

Keene, Suzanne. 1994. “Real-Time Survival Rates for Treatments of Archaeological Iron.” In *Ancient & Historic Metals: Conservation and Scientific Research. Proceedings of a Symposium Organized by the J. Paul Getty Museum and the Getty Conservation Institute, November 1991,* edited by David A. Scott, Jerry Podany, and Brian B. Considine. Marina del Rey, CA: Getty Publications. http://www.getty.edu/conservation/publications_resources/pdf_publications/ancientmetals.html. Accessed April 18, 2017.

Kelsey, Jennifer L. 1996. *Methods in Observational Epidemiology.* New York: Oxford University Press.

Kimberly, Arthur E., and Aelaide L. Emley. 1933. “A Study of the Deterioration of Book Papers in Libraries.” In *Bureau of Standards Miscellaneous Publications No. 140.* Washington, DC: U.S. Department of Commerce, Bureau of Standards. https://archive.org/details/studyofdeteriora-140kimb. Accessed April 18, 2017.

Koestler, R. J, P. Brimblecombe, D. Camuffo, W. S. Ginell, T. E. Graedel, P. Leavengood, J. Petushkova, M. Steiger, C. Urzi, V. Verges-Belmin, and T. Warscheid. 1994. “Group Report: How Do External Environmental Factors Accelerate Change?” In *Durability and Change: The Science, Responsibility, and Cost of Sustaining Cultural Heritage, December 6–11,* 1992. Hoboken, NJ: John Wiley.

Krzemien. Leszek, Michal Lukomski, Agnieszka Kijowska, and Bozena Mierzejewska. 2015. “Combining Digital Speckle Pattern Interferometry with Shearography in a New Instrument to Characterize Surface Delamination in Museum Artefacts.” *Journal of Cultural Heritage* 16 (4): 544–50.

Larsen, R. 1994. “A Review of the Results of the STEP Leather Project.” In *Environnement et conservation de l’écrit, de l’image et du son: Actes des Deuxièmes Journées Internationales d’Etudes de l’ARSAG, 16 au 20 mai 1994,* 48–55. Paris: Association pour la Recherche scientifique sur les arts graphiques.

Martin, J. W., J. W. Chin, and T. Nguyen. 2003. “Reciprocity Law Experiments in Polymeric Photodegradation: A Critical Review.” *Progress on Organic Coatings* 47: 292–311.

Melin, Charlotta Bylund, and Mattias Legner. 2014. “The Relationship between Heating Energy and Cumulative Damage to Painted Wood in Historic Churches.” *Journal of the Institute of Conservation* 37 (2): 94–109.

Michalski, Stefan. 2014. “The Power of History in the Analysis of Collection Vulnerabilities.” Paper presented at the ICOM-CC 17th Triennial Conference, “Building Strong Culture through Conservation,” 15–19 September 2014, Melbourne, Australia.

________. 2016. “Climate Guidelines for Heritage Collections: Where We Are in 2014 and How We Got Here.” In *Proceedings of the Smithsonian Institution Summit on the Museum Preservation Environment,* edited by Sarah Stauderman and William G. Tompkins, 7–32. http://opensi.si.edu/index.php/smithsonian/catalog/book/111. Accessed April 18, 2017.

Museum of Fine Arts, Boston. 2016. The Conservation and Art Materials Encyclopedia Online (CAMEO). http://cameo.mfa.org/wiki/Copper_number. Accessed June 20, 2017.

Popper, Karl R. 1959. *The Logic of Scientific Discovery*. New York: Basic Books.

Reedy, Terry J., and Chandra I. Reedy. 1988. *Statistical Analysis in Art Conservation Research, Research in Conservation.* Marina del Rey, CA: Getty Conservation Institute. https://www.getty.edu/conservation/publications_resources/pdf_publications/pdf/statistics.pdf. Accessed April 18, 2017.

Rohdin, Patrik, Mariusz Dalewski, and Bahram Moshfegh. 2012. “Using an Epidemiological Approach as a Supporting Tool for Energy Auditing of Culturally and Historically Valuable Buildings.” In *Postprints from the Conference Energy Efficiency in Historic Buildings, Visby, February 9–11, 2011,* edited by Tor Broström and Lisa Nilsen, 164–74. Visby, Sweden: Gotland University Press.

Shanteau, James. 2001. “What Does It Mean When Experts Disagree?” In *Linking Expertise and Naturalistic Decision Making,* edited by E. Salas and G. Klein. Mahwah, NJ: Lawrence Erlbaum Associates.

Staniforth, Sarah. 2014. “Environmental Conditions for the Safeguarding of Collections: Future Trends.” *Studies in Conservation* 59 (4): 213–17.

Strlic, M., D. Thickett, J. Taylor, and M. Cassar. 2013. “Damage Functions in Heritage Science.” *Studies in Conservation* 58 (2): 80–87.

Strojecki, Marcin, Michal Lukomski, Leszek Krzemien, Joanna Sobczyk, and Lukasz Bratasz. 2014. “Acoustic Emission Monitoring of an Eighteenth-Century Wardrobe to Support a Strategy for Indoor Climate Management.” *Studies in Conservation* 59 (4): 225–32.

Suenson-Taylor, Kirsten, Dean Sully, and Clive Orton. 1999. “Data in Conservation: The Missing Link in the Process.” *Studies in Conservation* 44: 184–94.

Sully, Dean, and Kirsten Suenson-Taylor. 1996. “A Condition Survey of Glycerol Treated Freeze-Dried Leather in Long Term Storage.” In *Preprints of the Contributions to the Copenhagen Congress, 26–30 August 1996: Archaeological Conservation and Its Consequences,* edited by Ashok Roy and Perry Smith, 177–81. London: International Institute for Conservation of Historic and Artistic Works.

________. 1999. “An Interventive Study of Glycerol Treated Freeze-Dried Leather.” In *Proceedings of the 7th ICOM-CC Working Group on Wet Organic Archaeological Materials conference = Actes de la 7ème Conférence du Group de travail Matériaux archéologiques organiques humides de l’ICOM-CC: ICOM-CC WOAM 98: Grenoble, France, 1998,* edited by Céline Bonnot-Diconne, Xavier Hiron, Quôc Khôi Tran, Per Hoffmann, 224–31. Grenoble: Atelier régional de conservation-Nucléart, 1999.

Taylor, Joel. 2013. “Causes and Extent of Variation in Collection Condition Survey Data.” *Studies in Conservation* 58 (2): 95–106.

________. 2014. “The Impact of Assessment Guides on the Reliability of Collection Condition Surveys.” Paper presented at the ICOM-CC 17th Triennial Conference, “Building Strong Culture through Conservation,” 15–19 September 2014, Melbourne, Australia.

Taylor, Joel, and Siobhan Stevenson. 1999. “Investigating Subjectivity within Collection Condition Surveys.” *Museum Management and Curatorship* 18 (1): 19–42.

Turgoose, S. 1985. “The Corrosion of Archaeological Iron during Burial and Treatment.” *Studies in Conservation* 30: 13–18.

PART 2

Summary of the Experts Meeting

Foekje Boersma

It is acknowledged that the issue of collection and material environmental requirements is complex, and conservators/conservation scientists should actively seek to explain and unpack these complexities.

IIC International Institute for Conservation of Historic and Artistic Works and
ICOM-CC International Council of Museums–Committee for Conservation 2014

The meeting was held in June 2015 at the Rothschild Foundation at Windmill Hill on the Waddesdon Estate, a new, purpose-built study center that houses the Rochester family archives. It convened a group of researchers active in the study of the behavior of materials when exposed to fluctuating climatic conditions and conservation professionals working with collections to explore possible ways in which epidemiological approaches could help in the investigation of the causal relationships between objects' mechanical damage and their environment (See Appendix 2, List of Participants). The meeting was moderated by Sarah Staniforth, president of the International Institute for Conservation of Historic and Artistic Works (IIC).

Over the course of two days, the group shared experiences and discussed the idea of applying epidemiological approaches to studying the state of preservation of collections. The objectives of the meeting were to identify the methodology and feasibility of an epidemiology study, to discuss its scope, and to identify areas for potential subsequent collaboration.

One of the main outcomes of the meeting was the adaptation of a common model used in epidemiology to rank the quality of evidence for use in conservation science. The visual representation in figure 11 above shows how our research and collection data may be organized in a pyramid. This quality of evidence pyramid builds from unfiltered background information up through basic study designs into critically appraised articles, validated predictive models including web-based tools, and, finally, systematic reviews including metastudies. These all build on the evidence to demonstrate, and often prove, causality.

The layers in the pyramid were also found to represent different methodologies for gathering data. At the base of the pyramid are environmental history, provenance data, and expert opinion, the building blocks that inform everything above. In the next layer up there is important data that would inform the mechanical damage question. Condition reports, cohort studies, and laboratory studies, including experimental models, constitute the next layer up, where more engineering-level data on historic materials behavior would be collected to ultimately inform the predictive or decision-assisting models. At the top are systematic reviews and metastudies.

Currently the majority of our research and documentation is located within the layers of unfiltered information. In order to make better use of this information, data collected should be in a form that allows for more effective comparison and exchange. Therefore, there is a need to establish protocols and a common language. These should address such research issues as how to reliably and adequately characterize the climatic conditions surrounding an object and protocols for applying acoustic emission as a technique for the direct tracing of physical change in historical materials at the microscale. Protocols and common language should aim toward some form of standardization of condition reporting, which would make comparing the recorded information easier. A comprehensive list should be created to allow for the collection of data that is more comparative, a list that includes visual descriptors, object characteristics, guidelines for measurements, and imaging techniques (photography).

In fact, it was suggested that the conservation field may apply this same quality of evidence approach to all projects currently under way or in the process of theoretical design. Assuming that a more rational set of "best practices" for sustainable museum environmental conditions rests on the strongest evidentiary foundation that we can provide, a network of research could build to that objective and, more specifically, identify where lacunae in the structure may be found.

It was suggested that is important to look at existing studies to inform new ones. The above-mentioned lack of protocols for data collection and a comprehensive list of observable phenomena was seen as a major gap. A critical review of the existing literature is also necessary. It was recognized that although critically appraised articles are published appropriately in the most widely recognized scientific journals, the language becomes a barrier to the field when it is digested in a forum like Studies in Conservation. Education and dissemination should help vernacularize the literature.

Suggestions for an epidemiological study were discussed, and it was established that the focus would be the identification of climate-induced mechanical damage to susceptible materials. Mechanical damage is the result of loading conditions, which can be mechanical, thermal, or hygral in nature. Since organic and hygroscopic materials are the most affected by climatic fluctuations, objects constructed from these materials were chosen as focus of the study. The research can include objects in a wide range of environments, from collections in libraries and archives to churches, historic buildings, or museums. Objects without damage would be included, especially those exposed to one or more climatic events that theoretically would have led to damage yet have not exhibited damage. Loss of value versus degree of negative change, acceptable rates of damage, and life expectancy are recognized as important considerations for management, but these concerns are outside the scope of this study, which is intended to remove as many variables as possible.

The advantages and disadvantages of retrospective versus prospective studies were discussed. It was acknowledged that retrospective reconstructions of conditions are challenging. Because objects have been moved, they have experienced different conditions and states. In prospective studies, one has control over data gathering, so this may be a more productive approach for the pilot study. However, retrospective studies should not be dismissed. Evidence of past causality or its absence exists even if the interpretation is not without considerable uncertainty.

The strength of epidemiology is that one has a large sample size, which allows the phenomenon of interest to be consistently identifiable among an array of situational variables. In terms of sample size, a correlation between the layers of the pyramid (see fig. 11) and the sensitivity of the observational technique was identified. The large groups of observed objects that underpin the pyramid correlate to categorical observations. Increasing the sensitivity of the technique used to monitor objects corresponds to the next levels in the pyramid (case studies, travel and exhibition reports, cohort studies, technical analysis, and lab studies), reducing the sample size (fig. 13). This is seen as a way to make studies more feasible.

Several areas of interest for informal collaboration were identified, with the expectation that these areas can inform one another. Some participants work with large collections and gather information through observations, which range from simple condition reports to photography. Others work in laboratories on detailed observations of individual objects, especially where research concerns fracture mechanisms using techniques like acoustic emission and strain gauges. Acoustic emission was identified as a highly sensitive technique that can now be applied in the field. Although interpretation of the data is still difficult, this emerging group of experts is eager to exchange data. Measurements using strain gauges will perhaps not result in information on damage (instead they give information on the overall deformation of an object), but this is necessary for the validation of models. Digital image correlation (DIC) appears to be a very cost-effective way to do long-term monitoring and may replace strain gauges.

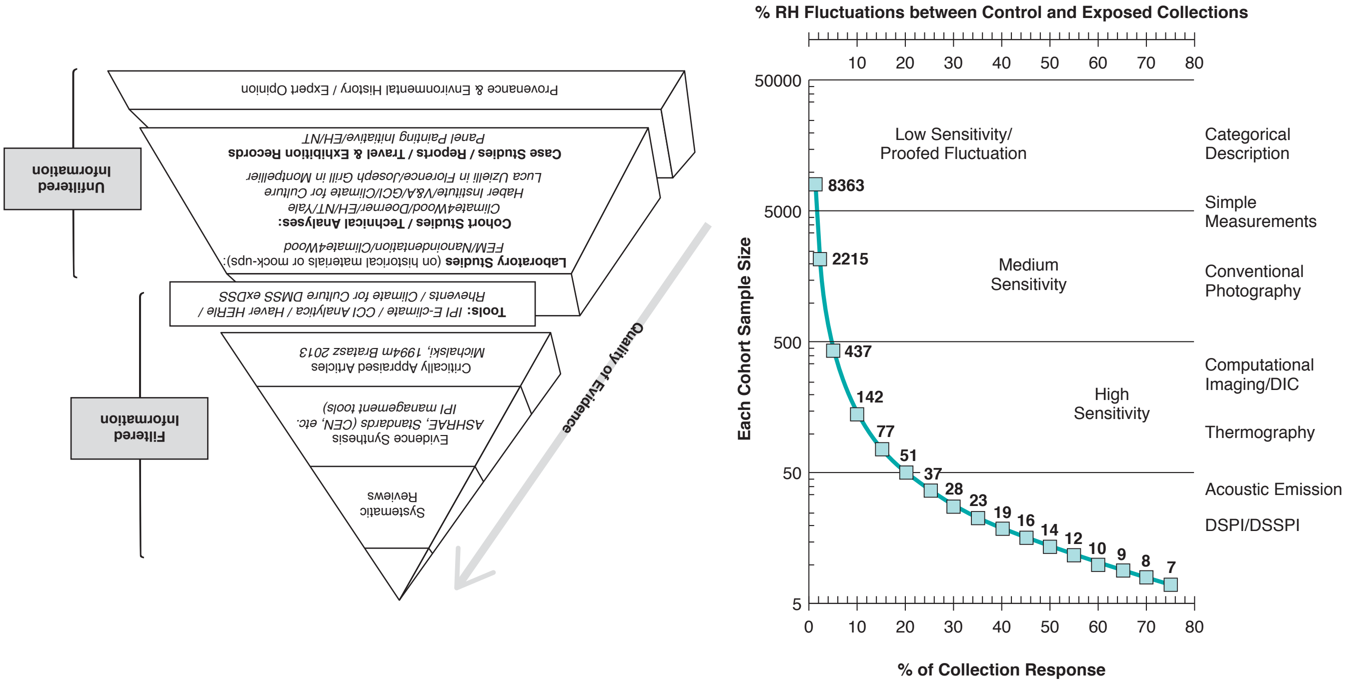


FIGURE 13.
The Quality of Evidence pyramid upside down (left) shows alignment with layers of information related to different sample sizes based on the sensitivity of the research technology (right).

A third group looks into big data topics: data merging and the vetting and analysis of that merged data. This also includes communications and web tools, making the research more accessible and applicable for management purposes. Several tools have been used or are under development (e.g., Analytica, Haber's risk assessment tool, IPI's e-climate, V&A RHevents).

Considering the sample sizes required, meeting participants agreed on the importance of working in collaboration with one another and with other colleagues active in this area elsewhere in the world, so that collectively we can have a much greater impact. Although specific formal collaborative projects were not identified, the participants expressed the wish to work together on an informal level, sharing information and assisting with data analysis and interpretation.

Epidemiological approaches may have been applied to individual projects in the past, but it is understood that our field as a whole is just recognizing the importance of synthesizing research on a larger scale, which requires more refined thinking. It will be necessary to conduct preliminary, pilot studies to determine what is feasible and to develop a methodological approach.

Pilot Cohort Study Design
(Retrospective/Prospective)
A
B
Cohort 1 (n similar objects) Historical Property 1
Cohort 2 (n similar objects) Historical Property 2
Existing Research
Methods
Observational
Photographic
Instrumental
Roman Kozlowski
Jim Reilly

FIGURE 14.
Jim Druzik, GCI senior scientist, introduces the concept of epidemiology at the meeting. Photo: Foekje Boersma, GCI (J. Paul Getty Trust)

Selected Comments by Participants

Lukasz Bratasz, head of the Sustainable Conservation Lab, Institute for the Preservation of Cultural Heritage, Yale University: “We all work, more or less, in this field of environmental impact on cultural heritage, especially on physical damages, but what was missing somehow was this nomenclature, the different understanding of biases, the criteria for epidemiology and so on.”

Nigel Blades, preventive conservation adviser (environment), National Trust (U.K.): “It is a real step forward to create a synthesis of the subject matter. There is clearly a lot of information, a lot of research, and a lot of knowledge residing in different areas and with different research groups. This meeting has provided a great opportunity to bring these different aspects together for discussion and to begin thinking about the synergies and how it can all come together in the future.”

Boris Pretzel, head of science, Victoria & Albert Museum, reflecting on how an epidemiological approach could strengthen the evidence on which climatic conditions are safe: “It would be really nice that, at the end of a few years, we have a higher level of certainty about the advice we’re giving. I think individually we often are quite convinced that we’re doing the right thing, but we need to know more, and this is a great way of going about it.”

APPENDIX 1

Glossary

Analytical epidemiology Studies that test causal relationships and make inferences fit into this category. This includes cohort and case-control studies but may also include randomized control trials or any other study that quantifies risk factors.

Base rate neglect The intuitive tendency to ignore the probability of some event's true chances of occurrence in the face of a specific example. This error stands out when a given test is true whenever a certain condition exists but will also render a false positive for a significant number of tests when it is not true. If the condition is rare, the probability of a false test is ignored or strongly deemphasized.

Bias Systematic attitudes or heuristics based on intuitive reactions rather than rational analysis. See page 14 of this document for examples.

Black swans The term given to highly unlikely occurrences that can falsify or have an impact on hypotheses. The concept of a black swan was introduced by Karl Popper (Popper 1959). Popper's assertion that falsifying information was more powerful and efficient was described by reviewing the statement, "All swans are white." Looking for white swans is an infinite task that could never conclusively prove the statement. Seeing one black swan would have an instant impact. "Falsifiability" is a Popperian term fundamental to hypothesis testing. A statement that is not falsifiable cannot be technically proven because regardless of the number of outcomes observed the possibility always exists that a contrary result can occur. However, a correlation or hypothesis can be made falsifiable when the search for a contrary result is first found.

Case-control study Involves two groups: one group having a disease, condition, or response, called "the cases"; and another group without a disease, condition, or response, called "the controls." A case-control study determines whether or not each individual in the two study groups has been exposed to some agent thought to cause the disease or condition. If the causal link is true, the prevalence of exposure would be higher for the cases than the controls.

Coefficient of determination (r^2) This statistic represents the proportion of the dependent variable (damage) that is predictable from the independent variable (environmental condition); 1 means perfect correlation, 0 means no correlation. ((r) is the correlation coefficient, a measure of the degree two variables indicate that they have a linear relationship to each other. Values range from +1 to –1. Lack of correlation does not mean a lack of relation, only a lack of a linear relation on the scale used to display it.)

Cohort study Involves two or more groups exposed to differing environmental conditions where the incidence of adverse effects could be shown to be causatively higher in the environment suspected of posing a greater risk.

Confounding variable A hidden or unrecognized variable that influences a result. A confounding variable is any variable that changes the result that was not accounted for.

Conservation heating The use of heating to lower relative humidity to an acceptable level by means of humidistatic control.

Copper number "A measured value used to determine the condition and stability of cellulose. The copper number is determined as the amount of copper reduced from the cupric to the cuprous state by 100 grams of cellulose pulp. A high copper number indicates that the cellulose is not pure and may have been degraded by bleaching. A low copper number indicates that the cellulose is not degraded. This test does not account for the presence of lignin" (Museum of Fine Arts, Boston 2016).

Cross-over study A study in which the respective environments or treatment options of the control and exposure groups are switched. Members are usually randomly ordered, and if unbiased or if the changes are permanent, this methodology can reduce experimental noise and sample size. Such a study is not an effective strategy for collections.

Cross-sectional studies Studies in which periodic assessments of all members of a group are conducted in order to determine changes in any of the parameters being tracked. These are also carried out as longitudinal studies and are usually prospective.

Descriptive epidemiology A survey or broad assessment that seeks to identify the distribution of a given condition but does not attempt to apply causal relationships to its frequency or distribution.

Dose-response function The relationship between the exposure and the response. The exposure may be a pathogen, drug, environmental condition or any stressor, which may be complex and nonlinear involving latent variables or direct and linear. Strlic et al. (2013: 80–87) point out that dose-response functions may also be called "change functions" to clearly decouple them from other functions that embrace cultural values. In this manner, change functions can be interpreted through value functions to understand overall damage.

Ecological fallacy; ecological inference fallacy A logical fallacy in the interpretation of statistical data where inferences about the association of individual and group variables (or among several group variables) show correlations that do not exist, for example, the correlation between number of observed storks and population growth or deaths by drowning each year and the release of Nicholas Cage movies. These are also called "nonsense correlations."

Epidemiology The study of the distribution of disease or another adverse condition within a population and the factors that influence this distribution.

Heuristic rules Often referred to as "rules of thumb," or methods that are derived from practical experience. In very specific instances they may have high reliability and be very efficient ways to process information, but generalized too broadly, they are often erroneous or biased.

High-validity environments. *See* Zero-validity environments.

Hypothesis testing A theoretical technique that begins by assuming that no statistically relevant relationship exists between a set of data and an effect. This is termed the null hypothesis. Subsequent hypotheses that posit a relationship are tested against the null hypothesis and a statistical value is derived called the p-value. Generally a p-value of 0.05 or lower is expected as a trigger to reject the null hypothesis as false and accept the alternative hypothesis as true.

Incidence The rate of acquiring a new disease or condition in a specified period of time for a population that is at risk.

Life tables The same as actuarial tables. They express the number, or pattern. of the population that has died, acquired a disease, or deteriorated to some endpoint within the target population over time-specific periods. Mortality and survival probabilities are determined for each period.

p-hacking The abuse of the p-value metric in hypothesis testing by collecting only data that support $p<0.05$, or the exclusion of member data of the study to achieve $p<0.05$. Another example is to collect and analyze many hypotheses until the researcher finds one, regardless of likelihood, that gives $p<0.05$, use covariates or other transformation to get $p<0.05$.

p-value A statistical test that returns the likelihood of a false positive hypothesis test, or one in 20 that the observed explanation for an apparent association in a dataset is incorrect. The value 0.05 is often chosen for scientific experimental analysis.

Prospective and retrospective studies Prospective studies begin at the present time. Their strength resides in the fact that all documentation can be designed from scratch and its reliability is more easily ensured by checks and balances. Retrospective studies are historical, with the end point being the present time. They rest on the documented record and draw their strengths or weaknesses from the completeness of those records.

Nominal, ordinal, or ratio variables Nominal variables: boxes or similar listings of characteristics. Ordinal variables: numeric or descriptive scales. Ratio variables express a condition as a percentage or ratio.

Null hypothesis A hypothesis that two or more variables connected to a hypothesis have no association to each other, or that two populations do not statistically differ in the area of study.

Prevalence The number of people or objects having a given disease or condition divided by the size of the population at any one time.

Quality of evidence (QOE) A term that describes both the *strength of evidence* (SOE), or the degree of positive evidence from a single study showing a statistical benefit, and the *weight of evidence* (WOE), the total positive and negative benefits integrated over all studies of a phenomenon.

Randomized control trial (RCT) A study in which test subjects are randomly assigned to a control or a test group. The test group is given some treatment or exposure and the control group is not. The effect of the treatment or exposure is then assessed after an appropriate period of time. The results are generalized to a larger population, but if the test subjects are not representative the accuracy of the experiment is suspect.

Survival probability The ratio of members of a population that have reached some specified endpoint to the number that has not. The end point may be death, a predetermined state of deterioration, or the resumption of some property, condition, or effect after treatment.

Wicked environments Professional environments where decisions have to be made in which cues that should, in theory, trigger good decisions actually trigger poor ones. These are rare, but they do exist. Studies have shown that executive compensation predicated on rewarding good performance actually encourages the opposite.

Zero-validity environments Professional environments in which decision-making cues can be capricious and unpredictable. The quality of the decision outcomes is seldom better than randomly selected choices. Professional stock pickers have been shown to be no more reliable than nonprofessional daily readers of the business sections of newspapers. Similar situations exist for probation officers and clinical psychologists. This is in contrast to *high-validity environments* where decision-making cues are consistent and reliable, such as for engineers, scientists, surgeons, and firefighters.

APPENDIX 2

Participants

Name	Organization
Nigel Blades	National Trust
Foekje Boersma	Getty Conservation Institute
Lukasz Bratasz	Institute for the Preservation of Cultural Heritage (IPCH), Yale University
Kathleen Dardes	Getty Conservation Institute
Jim Druzik	Getty Conservation Institute
Paul van Duin	Rijksmuseum Amsterdam
Melanie Eibl	Doerner Institut
Ralf Kilian	Fraunhofer Institut
Roman Kozlowski	Jerzy Haber Institute
Tom Learner	Getty Conservation Institute
Katy Lithgow	National Trust
Michal Lukomski	Getty Conservation Institute
Stefan Michalski	Canadian Conservation Institute
Boris Pretzel	Victoria & Albert Museum
James Reilly	Image Permanence Institute, Rochester
Sarah Staniforth	Consultant and President at the International Institute for Conservation of Historic and Artistic Works (IIC)
Akke Suiker	Technical University Eindhoven (TUe)
Joel Taylor	Norwegian Institute for Cultural Heritage Research (NIKU)
David Thickett	English Heritage

Pippa Shirley, head of collections at Waddesdon Manor (Rothschild Collections), was our host.

FIGURE 15.

Meeting participants left to right, top row: Kathleen Dardes, GCI; Stefan Michalski, CCI; Joel Taylor (at the time of the meeting NIKU, now GCI); Tom Learner, GCI; Michal Lukomski, GCI; Lukasz Bratasz, Yale; James Reilly, IPI; Roman Kozlowski, Jerzy Haber Institute; Ralf Kilian, Fraunhofer Institut; bottom row: Sarah Staniforth, president IIC and meeting moderator; Jim Druzik, GCI; David Thickett, English Heritage; Foekje Boersma, GCI; Boris Pretzel, V&A; Akke Suiker, TUe; Paul van Duin, Rijksmuseum; Katy Lithgow, National Trust (U.K.); Melanie Eibl, Doerner Institut; Nigel Blades, National Trust (U.K.).

Photo: GCI (J. Paul Getty Trust).

The Getty Conservation Institute