Using Epidemiology to Validate Scientific Results for Complex Situations

David Thickett
English Heritage
London, UK
david.thickett@english-heritage.org.uk

Abstract

Science can help in some aspects of improving conservation practice. Laboratory experiments have elucidated many factors for preventive conservation of archaeological iron and copper alloy objects. However, these rely on simple model systems (powder mixtures) and objects are much more complex in nature. Epidemiological methods can be readily adapted to conservation problems. Analysing the results of two large criteria-anchored surveys has shown that the experimental results do represent a good portion of archaeological iron and copper alloy responses. The graphs generated can be used for other collections and to carry out similar surveys and analyse the results. The approach requires very large numbers if based on visual observation, but these can be reduced with more sensitive analytical methods. Oxygen depletion has been demonstrated to provide useful conclusions using a much smaller sample set.

Keywords

iron, copper alloy, surveys, epidemiology, oxygen depletion

Introduction

Burial profoundly changes metal artefacts and can significantly alter their response to the environment. It introduces reactive species, alters the structure and dramatically increases porosity and often the fine-scale porosity responsible for uptake of water vapour and pollution. As well as significant differences between objects from the same archaeological site, burial can cause great differences throughout an object.

Scientific methods can help understand the important factors in preventive conservation and treatment. However, this requires near-identical replicates for exposure or treatment. Three broad options exist:

- the use of low-value objects, such as nails;
- the use of replicates, which are particularly difficult to produce for archaeological metals;
- exposure of objects only to the same conditions they will be exposed to in display or storage, which, given changes are slow, mainly requires very sensitive analytical techniques.

Two pieces of work on the required relative humidity (RH) and carbonyl pollutant conditions for archaeological iron and copper alloy have been published (Thickett and Odlyha 2013, Thickett 2016). These used iron/iron chloride and copper/copper chloride powder mixtures and measured the amount of akaganeite formed or oxygen depleted to quantify deterioration. Significant analytical evidence exists that archaeological copper does indeed contain copper chloride (Scott 1990) and iron chloride was analysed in archaeological iron (Thickett 2012). Whilst this research provides powerful evidence to inform preventive conservation strategies, there is a concern that simple model mixtures may not fully represent very complex objects. This has been addressed in two ways:

- Two surveys were undertaken of almost all of the archaeological iron and copper alloy objects on display at English Heritage sites and objects in store from excavations at those sites.
- Measurements were taken of oxygen depletion rates from a large number of real objects and compared with the RH measured in other studies on powder mixtures.

Several important criteria for object response studies have already been considered within the field of epidemiology, such as sample size, appropriate statistics, experimental design, quality of evidence and potential bias (Kelsey et al. 1996, Fleis et al. 2003, Dean et al. 2021). One useful concept in epidemiology is the hierarchy of methods, which indicates the reliability of different types

of evidence, as summarized in Table 1. Randomised controlled studies (comparing two cohorts – groups) are considered the strongest of the unfiltered (experimental) methods. So-called filtered information is considered to be higher quality evidence in epidemiology. However, this does not exist at the moment in this field. There are certainly critical knowledge gaps in preventive conservation for a number of materials (Thickett and Lankester 2012). The sample size for such unfiltered studies is essentially determined by the smallest change detectable. Figure 1 shows the required sample size in each cohort for differing percentages of the exposed object cohort with observed deterioration (Kelsey et al. 1996, Fleis et al. 2003). An alpha value of 0.05 (95% certainty that the two cohorts represent the population from which they are drawn) and a beta value of 0.2 (80% chance of avoiding a false positive result) were used to construct Figure 1. This followed the methodology described in detail in the OpenEpi website (Dean et al. 2021: OpenEpi, Version 3, open.source.calculator, SSCohort). Figure 1 was generated by feeding several hundred values into the calculator to produce a smooth line. The OpenEpi site not only provides extensive background information but also leads a user through all the necessary steps and decisions to undertake these kinds of studies.

Table 1. Hierarchy of epidemiological methods

Study type	Information type	Quality of evidence
Systematic reviews	Filtered	
Critically appraised topics	Filtered	
Critically appraised individual articles	Filtered	
Randomized controlled trials	Unfiltered	
Cohort studies	Unfiltered	
Case-controlled studies	Unfiltered	
Background information/expert opinion	Unfiltered	

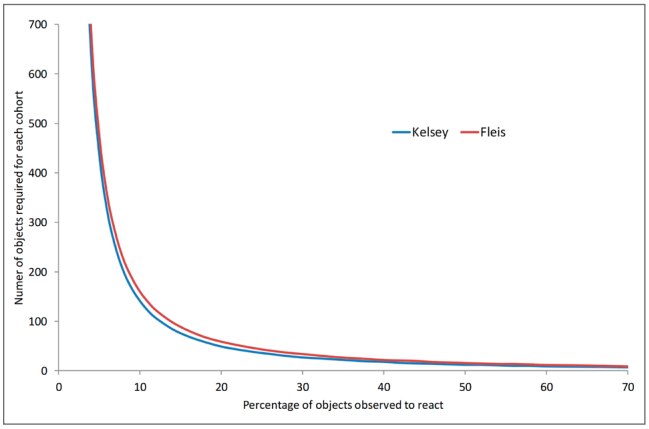

Figure 1. Number of objects required for cohort sizes for different observed reaction rates

The x scale is essentially the increased number of objects responding to the environment. The Fleis method produces a higher number of objects and, therefore, should be more robust. It uses continuity correction, a process to allow a continuous distribution (normal) to approximate a discrete distribution (binomial). Less sensitive detection, such as visual observation, will produce a smaller number of objects with observed changes. For example, for cracking wood, observation by eye may detect increases in crack length of perhaps 1 mm. More sensitive instrumental techniques, for example acoustic emission, has been shown to detect crack increases in wood in the order of microns, which means changes in more objects will be detected. This moves the experiment into the right-hand portion of the graph, requiring less objects in each cohort. Advances in instrumentation promise significant improvements over visual observation, hence higher percentages of objects with observable deterioration. The curves in Figure 1 can be used in other studies, provided the researchers are prepared to accept the alpha and beta value used.

The feasibility of cohort studies with visually assessed survey data of archaeological metals, supplemented by analysis, was investigated. In the first study, objects from the same archaeological site were grouped together. Different burial environments are known to affect reactivity significantly. Conservation treatments can have a significant effect. Although complete conservation records from excavation to the present day were not available, objects from the same site are more likely to have had similar treatments than those from different sites. The first study investigated archaeological iron. Cohorts were formed from objects kept at different RH levels (5 or 10% bands of maximum RH) and under low (< 1000) and high (> 1001 μgm^{-3}) concentrations of acetic acid in a series of showcases.

The complex nature of archaeological iron and copper artefacts makes measuring deterioration rate difficult via normal analytical techniques, with several complex layers of corrosion needing to be identified and quantified. In many instances oxygen reduction is considered to be the main cathodic reaction. If it is also the rate limiting reaction, then measuring oxygen consumption would indicate the corrosion rate of iron or copper. Enclosing an object in an impermeable container of fixed volume or measured volume, controlling the RH

and measuring the oxygen content of the air periodically allows measurement of the oxygen consumption rate (Thickett 2021).

Whilst the oxygen depletion technique can provide very valuable results, there are limits to the atmospheres that can be generated and maintained in a very well-sealed container. The author is unaware of any method to produce mixtures of room-type pollutants, nitrogen dioxide, ozone and sulfur dioxide that is suitable for the closed environments of this testing. Whilst enclosing room air provides an initial concentration, this will tend towards zero as the gas is consumed in reaction. Any pre-mixed atmospheres would rapidly lose pollution gases as they react with the objects, altering the concentrations. Additionally, heritage environments can be quite complex and fluctuate over wide ranges with time.

Methods

Surveys

Archaeological iron

For each of 31 sites, archaeological iron objects on display and in low RH storage were surveyed. Just over 1,600 objects from 121 showcases and 1,200 objects in store were investigated. The objects were assessed visually, using a criteria-anchored methodology (Sully and Suenson-Taylor 1996). Cracking, flaking and percentage powdering visible on the surface were assessed, according to four defined categories, as shown in Table 2. All objects were surveyed by the primary author and photographs of representative objects used to improve long-term consistency (Thickett and Odlyha 2013). Many of the displays had condition photographs taken at the time of installation. These were consulted to determine if the present damage had occurred during the display period.

The results were assessed to determine if enough objects had been surveyed to produce a statistically valid randomised controlled study (Kelsey et al. 1996, Fleis et al. 2003). The sample size for such studies is determined by the increase in proportion of the exposed sample cohort with the outcome over the unexposed sample cohort or, in this instance, the sample cohort exposed to different conditions. Samples of corrosion were taken from all objects and analysed by Fourier transform infrared (FTIR) spectroscopy (PerkinElmer 2000 FTIR with Amplif-IR diamond ATR using 32 scans and 4 cm^{-1} resolution) and some by X-ray diffraction (XRD) (Phillips 1830/1840).

As the room environments and showcases for the iron were highly variable, the display material had been exposed to a wide range of environmental conditions (RH and acetic acid concentrations). RH was measured with SmartReader SR002 loggers or Meaco radiotelemetry sensors with Rotronic HygroClip 2 probes. For each showcase, changing silica gel or the seasonal RH pattern leads to one or two maximum RH values per year. These values were recorded over several years. Acetic acid concentrations were determined using diffusion tubes analysed with a Dionex DX600 ion chromatograph and AS14 column (Gibson et al. 1997). One measurement (two replicate analyses if within 10% of each other and repeated if otherwise) was undertaken in each showcase during August or September. Experience has shown that these months give the highest acetic acid concentrations in most showcases in rooms in the UK without air conditioning. The RH results were initially considered in 5% bins (e.g. 35%–40%) of maximum RH measured for two sets of showcases, those above 1001 $\mu g/m^3$ acetic acid and those below 1000. If a maximum reading was not reached for half of the instances (each showcase had between 9 and 70 maximums), a lower maximum RH was used which met that criterion. The OpenEpi website was used to undertake t-tests (two-sided, 95% confidence

Table 2. Criteria used for surveys

	Iron			Copper alloy
Degree, score	**Corrosion visible on surface**	**Cracking**	**Flaking**	**Corrosion visible on surface**
None, 0				
Some, 1	1 or 2 spots, < 1% coverage	1 or 2 cracks, < 1 mm total length	1 or 2 flakes, < 1% coverage	1 or 2 spots, < 1% coverage
Medium, 2	3 to 10 spots, < 5% coverage	1 or 2 cracks, < 5 mm total length	3 to 10 flakes, < 5% coverage	3 to 10 spots, < 5% coverage
Heavy, 3	> 10 spots, > 5% coverage	Multiple cracks, > 5 mm total length	> 10 flakes, > 5% coverage	> 10 spots, > 5% coverage

interval) comparing adjacent RH bins. If significant, these were accepted; if not statistically significant, they were repeated with the next two bins combined (10%) and then with the next three (15%) until statistically significant results were obtained (Dean et al. 2021). Objects from the lowest display RH were compared to those in store. Stored iron is kept below 16% RH in pollution-free (certainly in terms of carboxylic acids) polypropylene boxes.

Archaeological copper alloys

A series of previously reported results for observed bronze disease on circa 3,800 Egyptian copper alloy artefacts were also re-evaluated in these terms (Thickett et al. 2008). The data was reprocessed with the statistical methods described. That survey, undertaken two decades previously, did not use a criteria-anchored method, which was under development at the time. Photographs of so-called 'index objects' were used to define the assigned category boundaries. For some deterioration phenomena, there are issues with visual identification under low magnification. Different corrosion products or materials can look similar and be mistaken for each other (Argyropoulos, pers. comm., 2007; Thickett and Pretzel 2010). Samples of all observed corrosion products were collected and analysed with FTIR spectroscopy (Nicolet 510PC) using a diamond cell and beam condenser.

A further similar survey to that for iron was undertaken on 1,200 objects on display in 84 showcases at 41 sites and 1,000 objects in store. A number of objects were assessed with oxygen depletion.

Oxygen depletion

The oxygen depletion rates of objects identified in the surveys as being representative of the types and degree of deterioration were measured. Most sites showed two groups of objects from previous tests (Thickett 2021): those that did not react, even at very high RH, and those that did. Only reactive objects were selected. Objects were selected that had been exposed to at least 55% RH. The objects were placed in variously sized sealed glass containers (Quickfit laboratory glassware, Bernardin or Bocal mason jars, corrosion jars and borosilicate glass jars with polypropylene screw tops blocked with aluminium foil) with glycerol solutions to control RH (Milner and Dalton 1953). The RH was measured in 10% of the containers with calibrated iButton (± 2%) temperature and RH data loggers. No excursions from the expected conditions were determined. Oxygen concentration was measured with a PreSens 4 oxygen meter with PreSens Sp-PSt3-NAU-D7-YOP self-adhesive oxygen spots through the glass container walls.

Tests were undertaken at 30, 35, 40, 45, 50 and 55% RH to investigate where the curve appears to be most complex in shape. A group of 60 archaeological iron objects were analysed at each RH value sequentially. A similar-sized group of archaeological copper artefacts were also analysed. Higher RH values of 60 and 65% were also used for the 54 objects that had previously been exposed to those RH values on display.

Results

Surveys

Archaeological iron

Data from a selection of the 121 showcases assessed are shown in Figure 2.

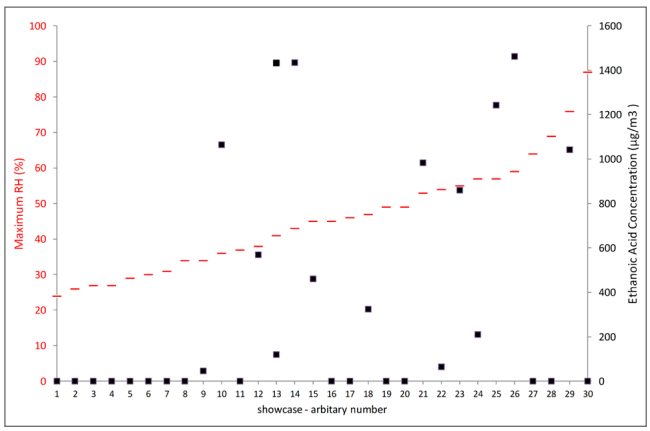

Figure 2. Maximum RH values and acetic acid concentrations in showcases (numbered 1–17)

This includes the maximum RH and acetic acid values and a spread of the values observed.

English Heritage has approximately 85% of its collection in store. For this approach, the maximum sample size was determined by the number of objects on display in similar conditions. The number of objects in dry storage is almost always significantly larger. Table 3 shows the object numbers on display and the calculated required cohort sizes for 11 selected sites.

Table 3. Selected cohort sizes

Site	Percentage reacting	Number required in cohort	Number of archaeological iron objects on display
1	11.18	139	286
2	12.31	122	276
3	25.00	46	104
4	26.67	41	39
5	26.88	41	92
6	27.08	41	73
7	28.88	37	90
8	29.44	37	38
9	28.57	37	72
10	29.74	34	64
11	66.67	11	14

The FTIR and XRD analyses showed three sites that contained a high proportion of corrosion products other than akaganeite (goethite and rozenite, iron sulfate). As the corrosion products are different, caution had to be exercised when considering these sites, and they were excluded from the analyses. The survey scores in a particular RH and acetic acid band were added to give the survey value. If a showcase had eight objects surveyed as 1, 1, 2, 2, 2, 3, 4, 4, the survey value was the sum (19 in this case). The survey score was normalised based on the number of objects and multiplied by 100 to give manageable numbers, i.e. 19/8, 237.5. All the showcases in the same acetic acid and maximum RH band were added together. The experimentally determined risk was then set to the survey score at the highest RH point (80%). Figure 3 shows the statistically significant points from the survey overlaid with the response line generated from powder experiments (partial results in Thickett and Odlyha 2013, full results in Thickett 2012). The points were first assessed for the number of objects in the cohort, then the corrosion products were determined and afterwards the statistical significance between different RH bands.

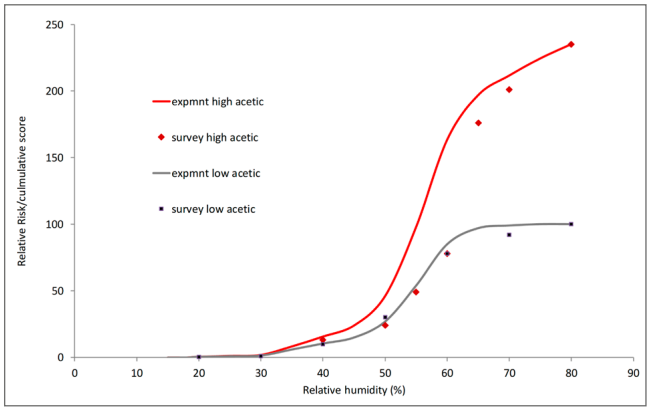

Figure 3. Experimentally derived reaction risk compared to cumulative scores from object assessments for low and high acetic acid environments

The grey and red lines are experimental data on powder mixtures from Thickett (2012). It was not possible to determine errors for the survey results. It would be possible to have multiple surveyors assessing the same objects, but as this would involve visits to 72 sites over a very wide geographic area, this was considered prohibitively expensive. Published research on surveys of archaeological iron objects indicated errors in the order of 10% (Leese and Bradley 1995). As can be seen, there is a very good correlation between the two data sets. Figure 3 also includes similar data for higher concentrations of acetic acid (red line and red diamonds).

Again, a good correlation is shown. There are no survey points below 40% RH, as these were all older wooden showcases with relatively high air exchange rates and even large amounts of silica gel, very low RHs cannot be maintained. Many of these cases have been replaced in the last decade, hence the smaller number of points.

Archaeological copper alloys

For the copper alloy artefacts, a much smaller percentage difference was observed, meaning larger sample sizes would be required. The number required for a significant response at the visually observed difference levels is shown in Figure 4.

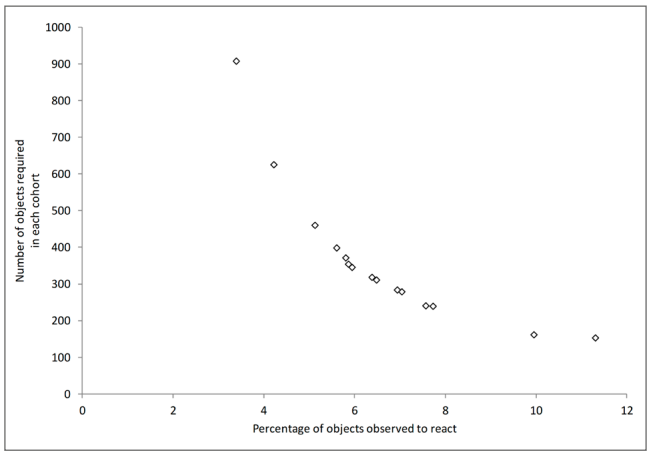

Figure 4. Number of objects required for cohorts for observed copper alloy degradation rates

This was generated by entering the percentage of non-category 0 objects from each site into the OpenEpi website. Taking the raw, visually assessed data, the 17 sites required cohort sizes of between 130 and 911. This is an example of the sample sizes required and no further statistical analyses were undertaken with this data. For

many collections, these numbers would rarely be on display. The effect of acetic acid (and formic acid and formaldehyde) is much smaller than for iron, so more showcases can be grouped together to provide larger numbers. The types of corrosion analysed for these objects are shown in Figure 5.

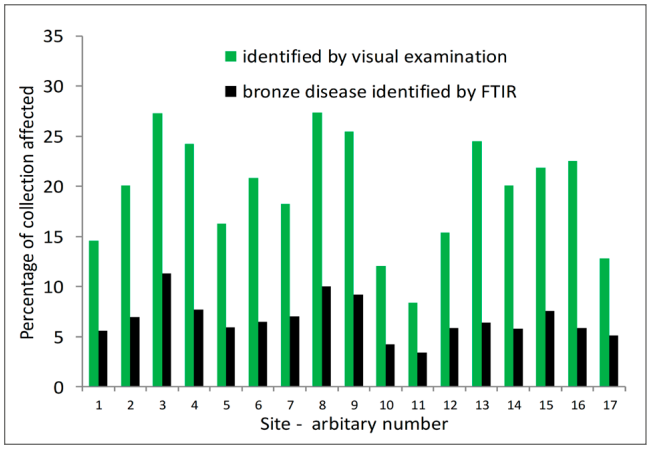

Figure 5. Copper alloy corrosion products determined by site

Many of the objects recorded as having bronze disease from visual examination (mainly spots of bright green corrosion) were found by FTIR analysis to have minerals other than atacamite or paratacamite (labelled as bronze disease in Figure 5) on their surfaces. Hence, the proportion reacting with actual bronze disease is lower and the required cohort size greater.

The significant results from the second copper alloy criteria-anchored survey are shown in Figure 6.

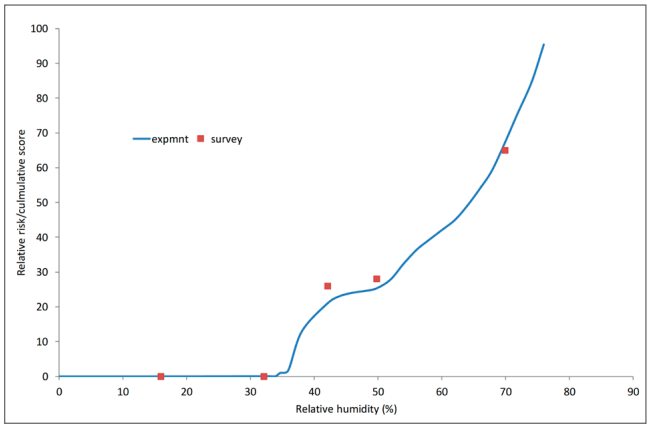

Figure 6. Experimentally derived reaction risk compared to cumulative scores from object assessments for copper alloys

The blue line is from experimental data on powder mixtures from Thickett (2012). Only five points were determined to be statistically significant. They follow the basic shape of the experimentally derived response curve, but because there is so little available data, it can only be considered an indication.

Oxygen depletion

Archaeological iron

Figure 7 shows the results of oxygen depletion measurements on 60 archaeological iron objects. Minimum, maximum and interquartile ranges are shown. The red bars are experimental powder results (same of the raw data can be seen in the grey line in Figure 3, but the graph is scaled to fit the survey results).

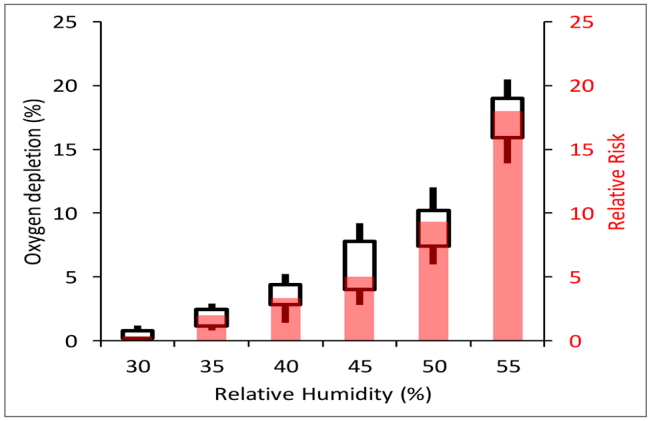

Figure 7. Oxygen depletion rates measured for 60 archaeological iron objects

There was a sizeable spread of data similar to the results reported by Watkinson et al. (2019). The ranking of each individual object compared to all those tested across the RH values tended to remain similar. It is reasonably clear that the bulk of the data follows the experimentally derived data.

Archaeological copper alloys

Figure 8 shows the results for the archaeological copper alloy objects.

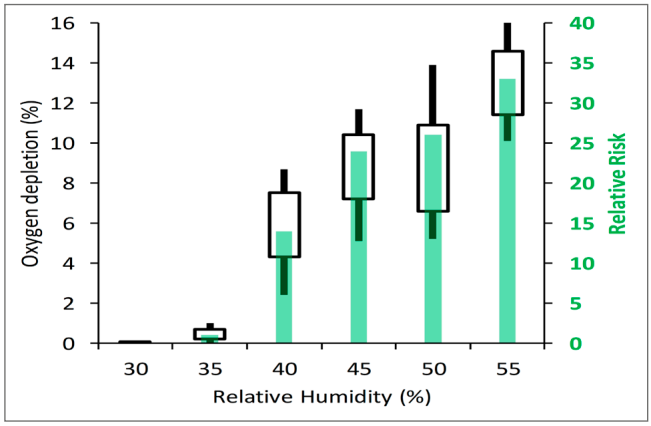

Figure 8. Oxygen depletion rates measured for 56 archaeological copper alloy objects

Again, within the range of responses, it appears the experimentally derived bars provide a reasonable fit for actual object response.

Discussion and conclusion

The surveys produced results in agreement with experimental powder studies, validating the adoption of their results and enhancing confidence in their use to improve preventive conservation. The iron powder research elucidated the impact on the reactions of RH, acetic and formic acid and formaldehyde concentration, temperature to a degree, the presence of copper (from associated burial), humic acid, goethite and akaganeites formed under different RH conditions, and the metal to chloride ratio. The impact of short periods of higher RH, such as when showcases or storage boxes are opened, and the influence of still and flowing air were also investigated. Many factors remain to be elucidated: the effect of the metallurgy of the object, the different forms of chloride present, differences in the reactivity and stability of the akaganeites present, the reactions generating goethite, and iron carbonate and iron sulfate. Whilst the selection of objects for display in each showcase is clearly not random, tests were undertaken that took into consideration object typology and burial depth using the diehard test suite (Florida State University 1995) for four sites. Over 98% of the p-values were between 0.025 and 0.975, indicating the randomness tests were passed at a 0.05 level.

The number of objects on display was found to be a limiting factor in several instances and especially with the less responsive copper alloy objects. The oxygen depletion measurements produced more finely graded results and could determine much smaller differences in response with smaller numbers of objects. Techniques from epidemiology proved extremely useful to the development of the study. Together, the methods provided confidence that the experimental corpus of results represents a good portion of archaeological iron and copper alloy responses. The methods described can readily be used by other researchers. The OpenEpi website guides users through the decisions needed. The cohort number requirements shown in Figure 1 give a general indication for required sample sizes for such studies. Previous work has shown that surveys of small subsets can give fairly robust estimates for percentage reaction rates for archaeological metal collections. The distribution of the results determines whether statistically significant differences can be detected between cohorts, and these numbers are always higher than those from the Fleis equation. Epidemiology also has potential as a primary research method and the advantage of considering the full complexity of real objects in real environments provided methods to accurately determine small changes are developed.

References

Argyropoulos, V. Technological Educational Institute of Athens, Greece, personal communication, 25 February 2007.

Dean, A.G., K.M. Sullivan, and M.M. Soe. OpenEpi: Open source epidemiologic statistics for public health, version 3.01, updated 2013/04/06. www.openepi.com/ (accessed 25 October 2021).

Fleis, J.L., B. Levin, and M.C. Paik. 2003. *Statistical methods for rates and proportions*, 3rd ed. New Jersey: John Wiley & Sons.

Florida State University. 1995. *The Marsaglia random number CDROM including the diehard battery of tests of randomness*. Tallahassee: Florida State University.

Gibson, L.T., B.G. Cooksey, D. Littlejohn, and N.H. Tennent. 1997. A diffusion tube sampler for the determination of acetic acid and formic acid vapours in museum cabinets. *Analytica Chimica Acta* 341: 11–19.

Kelsey, J.L., A.S. Whittemore, A.S. Evans, and W.D. Thompson. 1996. *Methods in observational epidemiology.* Oxford: Oxford Academic Press.

Leese, M.N. and S.M. Bradley. 1995. Conservation condition surveys at the British Museum. In *CAA94. Computer applications and quantitative methods in archaeology (BAR International Series 600)*, eds. J. Huggett and N. Ryan, 81–86. Oxford: Tempus Reparatum.

Milner, C.S. and N.N. Dalton. 1953. *Glycerol*, 269. New York: Reinhold Publishing Corporation.

Scott, D.A. 1990. Bronze disease: A review of some chemical problems and the role of relative humidity. *Journal of the American Institute for Conservation* 29(2):193–206.

Sully, D. and K. Suenson-Taylor. 1996. A condition survey of glycerol treated freeze-dried leather in long-term storage. In *Archaeological Conservation and its Consequences: Preprints of the Contributions to the Copenhagen Congress, 26–30 August 1996*, eds A. Roy and P. Smith, 177–81. London: International Institute for Conservation.

Thickett, D. 2012. Post excavation changes and preventive conservation of archaeological iron. PhD dissertation, University of London, UK. Available online at www.english-heritage.org.uk/siteassets/home/learn/conservation/collections-advice--guidance/thickettthesisfinalversion.pdf (accessed 25 October 2021).

Thickett, D. 2016. Critical relative humidity levels and carbonyl pollution concentrations for archaeological copper alloys. In *Metal 2016: Proceedings of the Interim Meeting of the ICOM-CC Metals Working Group, New Delhi, 26–30 September 2016*, eds. R. Menon, C. Chemello, and A. Pandya, 0401_24. New Delhi: International Council of Museums-Committee for Conservation (ICOM-CC) and Indira Gandhi National Centre for the Arts (IGNCA).

Thickett, D. 2021. Oxygen depletion testing of metals. *Heritage* 4(3): 2377–89.

Thickett, D. and B. Pretzel. 2010. Micro-spectroscopy: A powerful tool to understand deterioration. *E-Preservation Science* 7: 158–64.

Thickett, D. and P. Lankester. 2012. Critical knowledge gaps in environmental risk assessment and prioritizing research. *Collections* 8(4): 281–96.

Thickett, D. and M. Odlyha. 2013. The formation and transformation of akaganeite. In *Metal 2013: Proceedings of the Interim Meeting of the ICOM-CC Metals Working Group, Edinburgh, 16–19 September 2013*, eds. E. Hyslop, V. Gonzalez, L. Troalen, and L. Wilson, 103–109. Edinburgh: International Council of Museums-Committee for Conservation (ICOM-CC) and Historic Scotland.

Thickett, D., S. Lambarth, and P. Wyeth. 2008. Determining the stability and durability of archaeological materials. In *Art08: 9th International Conference on NDT of Art, Jerusalem, Israel, 25–30 May 2008*. www.ndt.net/article/art2008/papers/024Thickett.pdf (accessed 25 October 2021).

Thickett, D., V. Vilde, P. Lankester, and E. Richardson. 2017. Using science to assess and predict object response in historic house environments. In *Preventive Conservation in Historic Houses and Palace Museums: Assessment Methodologies and Applications. Conference Proceedings, Palace of Versailles and Trianon, 29 November–1 December 2017,* ed. N. Francaviglia, 258–70. Milan: Silvana Editoriale SpA.

Watkinson, D.E., M.B. Rimmer, and N.J. Emmerson. 2019. The influence of relative humidity and intrinsic chloride on post-excavation corrosion rates of archaeological wrought iron. *Studies in Conservation* 64(8): 456–71.

Author

David Thickett has a degree in natural sciences, a PhD in archaeological conservation and chemistry and worked for two years in industrial ceramics research. He joined the British Museum in 1990, specialising in preventive conservation and inorganic materials conservation research. He joined English Heritage in 2003 as a senior conservation scientist, mainly researching preventive conservation. Recent projects have focussed on historic house environments, acoustic emission, collections demography and epidemiology, non-destructive testing, microclimate frames and optical coherence tomography. He sits as a UK expert to the European Standards CEN/TC 346 (conservation standards) and is a directory board member of the Infrared and Raman Users Group (IRUG).