Radiocarbon, Vol 00, Nr 00, 2019, p 1–16 DOI:10.1017/RDC.2019.55
Selected Papers from the 23rd International Radiocarbon Conference, Trondheim, Norway, 17–22 June, 2018

CONFESSIONS OF A SERIAL POLYGAMIST: THE REALITY OF RADIOCARBON REPRODUCIBILITY IN ARCHAEOLOGICAL SAMPLES

Alex Bayliss • Peter Marshall*

Historic England, Cannon Bridge House, 25 Dowgate Hill, London, EC4R 2YA, UK

ABSTRACT. Since 1993 Historic England (and its predecessor English Heritage) has commissioned 9074 radiocarbon (^{14}C) measurements on archaeological samples. Over 80% of these have been interpreted within formal Bayesian statistical models. The multiple strands of reinforcing evidence incorporated in these models provide precise chronologies that make stringent demands on the accuracy of the ^{14}C results included in the analysis. Inter-laboratory replication is consequently a routine part of model construction and validation. We report an analysis of replicate measurements on 1089 archaeological samples. It is clear that laboratory reproducibility accounts for only part of the observed variation. The type of material dated is also critical to the reproducibility of measurements, with some sample types proving particularly problematic.

KEYWORDS: AMS dating, ^{14}C dating.

INTRODUCTION

Bayesian statistics have been employed for scientific dating programs funded by Historic England, and its predecessor English Heritage, since 1993 when the release of the first version of OxCal enabled chronological modeling to be undertaken routinely by archaeologists (Bronk Ramsey 1994; Bayliss and Bronk Ramsey 2004; Bayliss 2009). Since then, 9074 radiocarbon (^{14}C) measurements have been commissioned on archaeological and palaeoenvironmental samples by the organization. Over 80% of these have been included in formal Bayesian chronological models, along with a similar number of dates that have been inherited from previous research.

Figure 1 illustrates the iterative approach to sample selection and chronological modeling that has been forged out of this body of practice in England over the past 25 years. Once the archaeological context of the situation has been considered and the problem to be addressed by the scientific dating program explicitly defined, the prior information available for inclusion in the model and a pool of potential samples are identified. Careful assessment of both the archaeological association between the sampled material and the problem at hand and the origin of the carbon in the sampled organism (e.g. reservoir effects, potential contamination) is required at this stage.

In practice there are three strands in creating an effective sampling strategy from these components:

- statistical simulation can estimate the number of samples that should be dated to have a realistic chance of addressing the objective of the dating program at the required resolution,
- the selected samples must be archaeologically representative as well as statistically viable,
- the risks of submitting a particular suite of samples for dating must be identified and, where possible, mitigated.

Replication forms one strategy in mitigating the risks inherent in ^{14}C dating any suite of samples.

*Corresponding author. Email: peter.marshall@historicengland.org.uk.

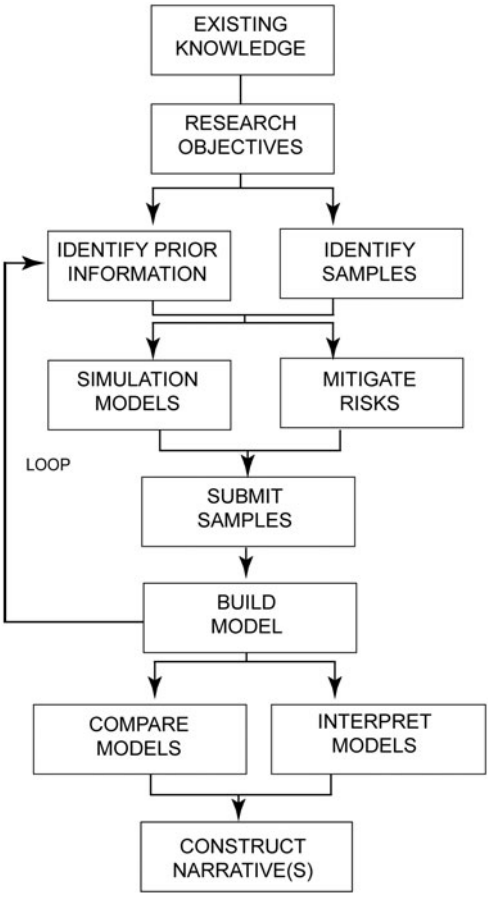

Figure 1 The Bayesian Process.

RISK AND REPLICATION IN THE BAYESIAN PROCESS

The risks of a sampling strategy for ^{14}C dating should first be mitigated by careful consideration of the archaeological and scientific strengths and weaknesses of each potential sample before its submission for dating (Bayliss et al. 2011). The perceived risks need to be balanced in selecting samples for dating. So, for example, single-entity dating of material that certainly derived from a single organism (Ashmore 1999) is a strategy which minimizes the risk that the submitted sample will contain reworked or intrusive material, thus returning a ^{14}C age that is the mean of the date of all items and the actual age of none of them. Selecting samples from a range of materials is a strategy which minimizes the risk that there will be a technical problem with dating a particular material type, for example where bone collagen is poorly preserved.

Once the samples reach the dating laboratory, a whole raft of internal laboratory quality assurance procedures come into play (e.g. McCormac et al. 2011). In addition to the use of the international reference material, Oxalic Acid II (Mann 1983), these include the use of background standards that are devoid of ^{14}C (sometimes embracing a range of materials

that are commonly dated, such as bone and wood, in addition to geological materials such as anthracite and calcite [e.g. van der Plicht et al. 2000: fig 5]). Laboratories also employ a range of secondary standards which are dated repeatedly, both as a check to identify when something may have gone wrong in processing a particular batch of samples and to determine over the long-term how the actual scatter of results compares with those expected on the basis of the quoted errors. Again, often these also cover the range of materials that are commonly dated (e.g. Brock et al. 2007: fig 3; Staff et al. 2014: fig 1).

^{14}C laboratories have a long-standing concern with the accuracy and reproducibility of their measurements (e.g. Willis et al. 1960) and, over the past 30 years, a series of formal international inter-comparison exercises has been undertaken (e.g. Scott et al. 2017). The results are used by the participating laboratories to identify and resolve technical problems with their sample processing and measurement systems, and also provide a suite of reference materials (many of which are used by laboratories as secondary standards). These studies provide spot-checks on the operational performance of the participating laboratories at the time the inter-comparison samples were analyzed. They do not measure consistent performance over a period of time, and so only the anonomized analysis of the reported results is usually published (Scott 2003: 151–248). Some laboratories do, however, choose to make their results public (e.g. McCormac et al. 2011: table 10 and figs 18–20), which can be useful when assessing the likely accuracy of legacy data or determining the likely scope of laboratory problems which are identified in retrospect. Recently, a number of smaller inter-comparison exercises, specialising in specific material types have also been undertaken (e.g. Naysmith et al. 2007).

The quality assurance protocols undertaken by ^{14}C laboratories, both individually and collectively, form the first strategy ensuring the accuracy of archaeological chronologies. These procedures can, and do, identify problems and allow them to be eliminated (e.g. Bronk Ramsey et al. 2002: 2). But there is always a risk that some issues will not be caught by these laboratory procedures.

The second strategy for reducing the risk of producing inaccurate dating is to test the accuracy of the ^{14}C dates once they have been obtained, both individually and as a group. There are a number of methods that we can use as a check on our results:

- the coherence of a suite of related ^{14}C dates—are there any clear outliers or misfits (see Bayliss et al. 2016: 56),
- the compatibility of a series of results with the relative chronological sequence known from archaeological information (such as stratigraphy; e.g. Bronk Ramsey 2009a, 2009b),
- the consistency of replicate results on the same or similar material (see Ward and Wilson 1978).

The first two methods come into play once our ^{14}C results have been reported, replicate samples, however, must be selected as part of the overall sampling strategy. Replication is neither scientific prurience nor an expensive luxury, but rather an essential element of any competent sampling strategy for ^{14}C dating.

There are two types of replicate measurement: multiple samples on different single-entities from the same context or feature, and replicate measurements on the same sample. The first mitigates the archaeological risk that the dated samples are residual, reworked or

intrusive; the second mitigates the scientific risks of dating particular materials. It is this second kind of replication that is considered in this paper.

In this paper we employ the approach to assessing the consistency of replicate measurements on the same object (Case I) described by Ward and Wilson (1978: 20–21). In interpreting the results of these tests, however, we are careful to take account of the assumptions of this method. These are that we have obtained statistically independent measurements, with normally distributed estimates of total error, on material that is certainly of the same ^{14}C age. Additionally, if we are to use the results to assess the reproducibility of a dataset, then the samples should have been selected for replication randomly.

The data considered here violate all these assumptions to a greater or lesser extent. No less than 486 of the 1089 replicate groups (45%) contain more than one measurement from the same laboratory, and so some data will share systematic factors. The assumption of normality within the quoted error is probably reasonable, however, as only 10 (< 1%) of the replicate groups are beyond two half-lives in age. Although each replicate group consists of measurements on a single sample, the ^{14}C age of that sample is not necessarily constant. For example, if different chemical fractions of that sample were selected for dating. Although the majority of samples were selected for replication randomly, in some cases repeat measurements were obtained in an effort to resolve a suspected problem with the initial result.

These issues are considered when assessing the consistency of replicate measurements on different types of samples below. The analysis does not explore the same problem in every case.

REPRODUCIBILITY IN PRACTICE

Figure 2 shows replicate groups in our dataset, categorized by their statistical consistency and material type. Overall 287 out of 1089 replicate groups (26%) are statistically inconsistent at the 5% significance level (Ward and Wilson 1978). The replicate groups contain 2,298 measurements, and so at least 12% of results report must lie more than 2σ from the true value. This compares to the 5% that would be expected simply on statistical grounds, although samples were not chosen for replication randomly and so this finding is not unexpected.

A small number of the replicate groups were obtained to investigate known technical problems (such as consolidant contamination) or to investigate the homogeneity of bulk samples, and about a fifth examine the reproducibility of measurements on different physical and chemical fractions of bulk sediment. The majority of replicates, however, were obtained on what were considered to be unproblematic samples. These were not randomly selected from all the samples submitted for dating, but rather selected judgmentally based on the perceived risk that the replication was intended to mitigate.

This led to a much higher proportion of replication for bone and antler samples, which require more complex pretreatment in the laboratory but whose archaeological taphonomy is often relatively unproblematic (many such samples were from articulating bone groups). In contrast, replicate measurements were obtained for relatively few samples of charred plant remains, as their processing in the laboratory is generally straightforward. Instead repeat measurements were obtained on different short-lived single entity samples from a feature in order to investigate the higher archaeological risk that the dated material was residual or

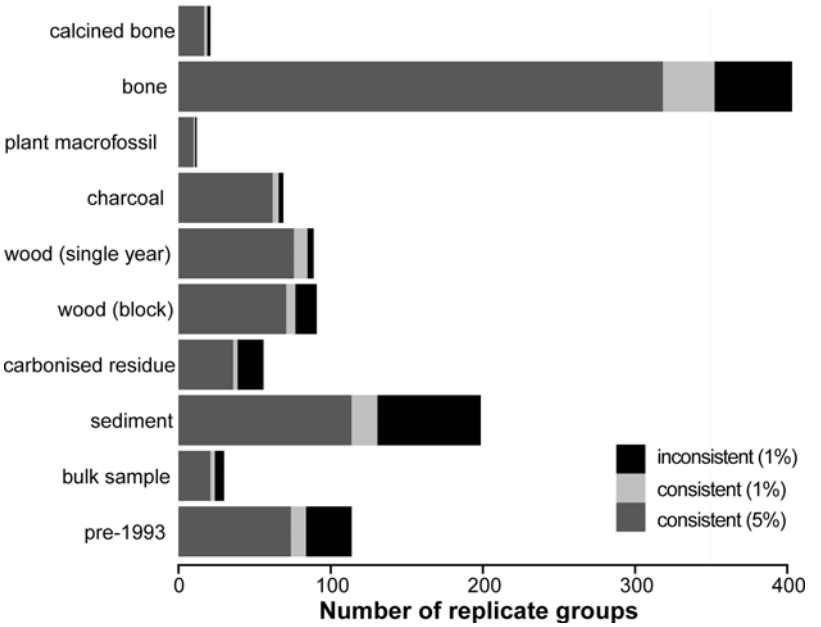

Figure 2 Bar chart showing the number of replicate groups of measurements on different material types which are statistically consistent at the 5% or 1% significance levels.

intrusive. For some sample-types, replication was constrained simply by the amount of material that was available (e.g. carbonized residues on pottery sherds).

Some materials are clearly more challenging: 85 of the 199 replicate groups on different chemical fractions of sediment (43%) are statistically inconsistent at the 5% significance level, for example, but only seven of the 69 groups on samples of carbonized plant material (10%).

Replicate Groups Containing Samples Dated Before 1993

A total of 114 replicate groups of measurements which contain at least one result produced before 1993 are available. Seventeen of these were undertaken to investigate known technical problems, 11 to investigate the homogeneity of bulk samples, and five on replicate chemical fractions of bulk sediment. Over half these replicate groups are statistically inconsistent at the 5% significance level (20 out of 33).

A further 81 replicate groups are available on samples that were re-measured either at the time of analysis because there was sufficient material available, or more recently by accelerator mass spectrometry (AMS) (sometimes as a check on the original measurement, but usually to obtain greater precision). Twenty of these replicate groups (33%) are statistically inconsistent at the 5% significance level (Figure 3a). Bone and antler (11 out of 25; 44%) and charcoal (8 out of 41; 20%) are clearly more problematic than other sample types. In the case of bone and antler this probably reflects the difficulty of providing sufficient material of this type for conventional dating in the 1970 s and 1980 s (and the pretreatment protocols that could thus be employed). The results on the charcoal are more likely to reflect variation in the proportion of old wood in the dated samples.

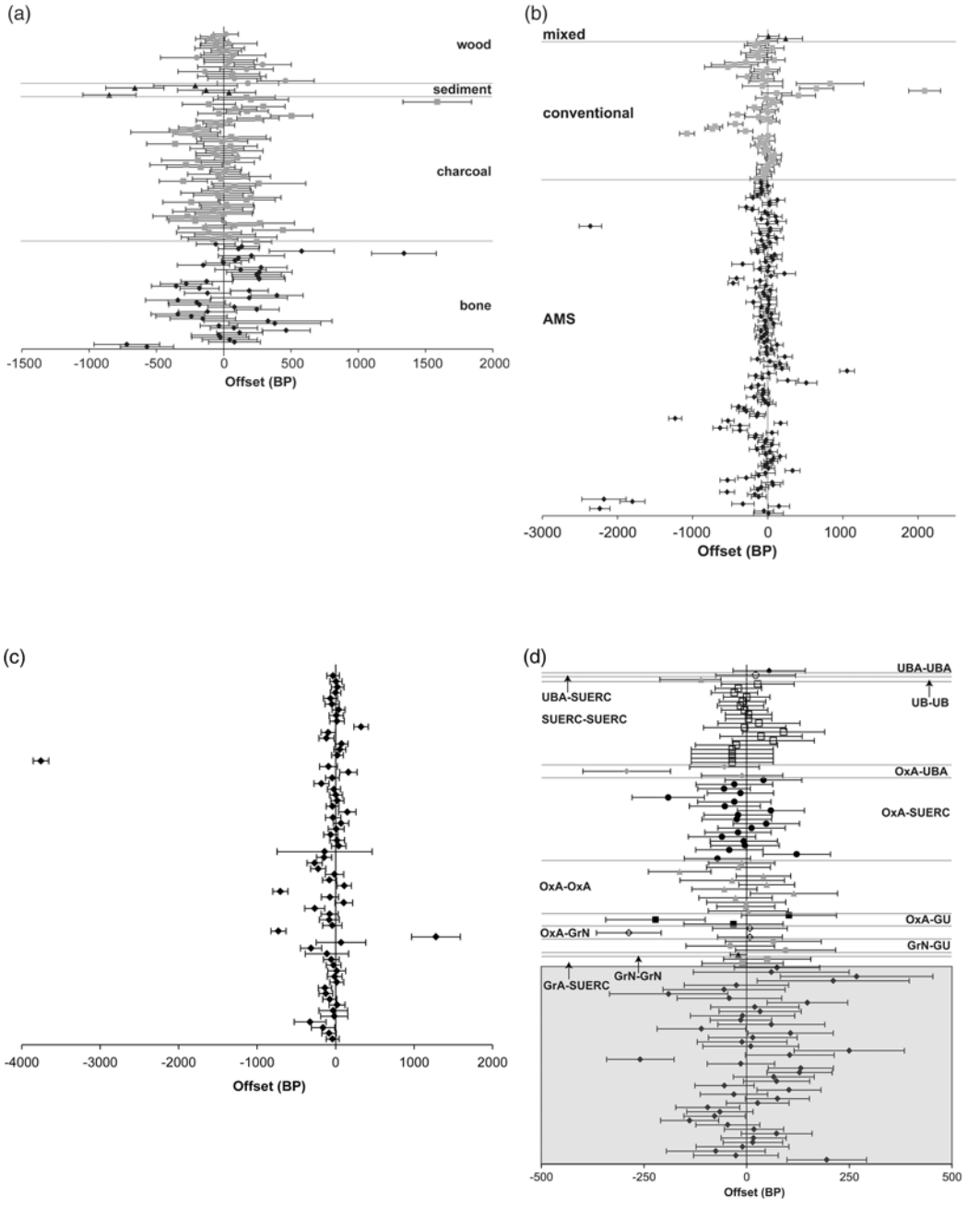

Figure 3a Offsets between pairs of replicate ^{14}C measurements, where at least one measurement in the group was made before 1993 (error bars at 2σ; if there are more than two measurements each is successively plotted against the first). Figure 3b Offsets between pairs of ^{14}C measurements on replicate "humic acid" and "humin" fractions of organic sediments (error bars at 2σ; if there are more than two measurements each is successively plotted against the first). Figure 3c Offsets between pairs of replicate ^{14}C measurements on carbonized residues on pottery sherds (error bars at 2σ; if there are more than two measurements each is successively plotted against the first). Figure 3d Offsets between pairs of replicate ^{14}C measurements on multi-ring blocks of waterlogged wood (error bars at 2σ; if there are more than two measurements each is successively plotted against the first).

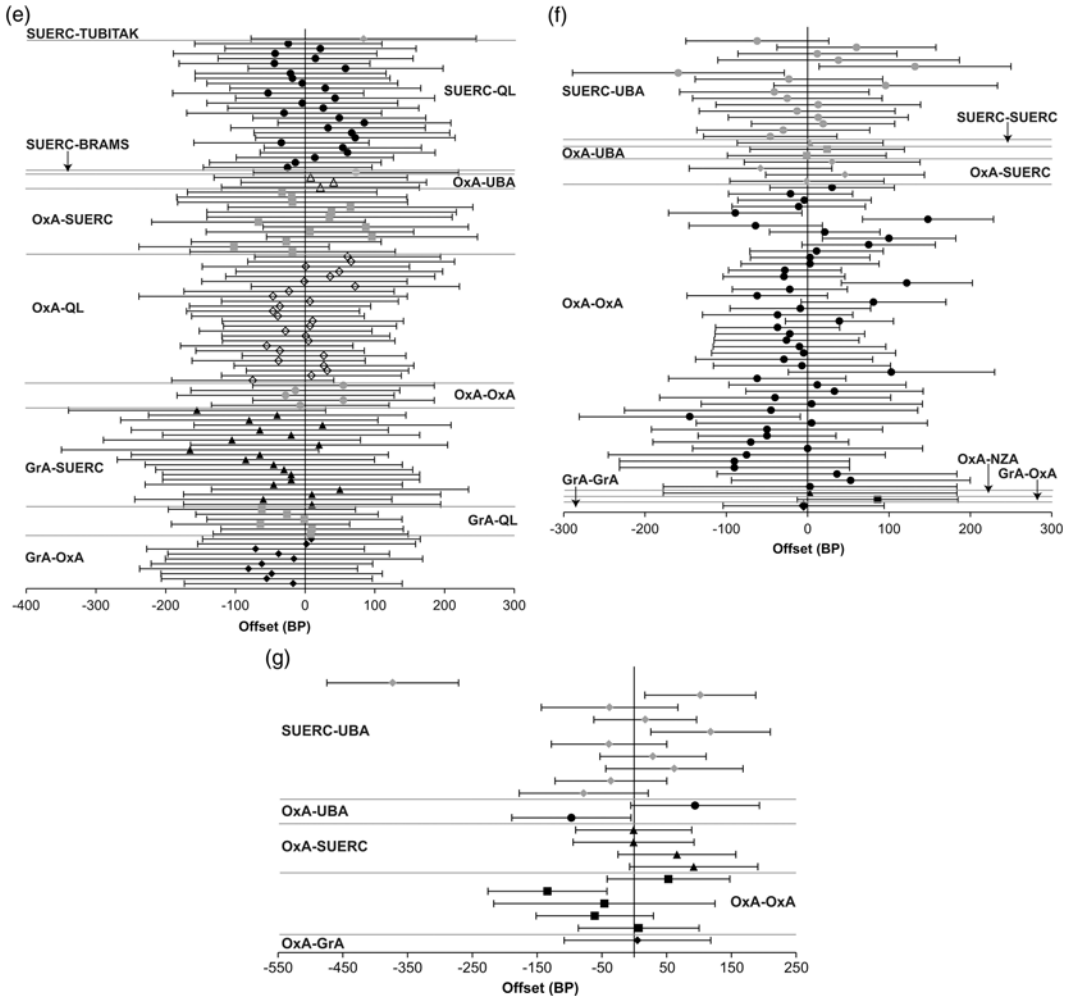

Figure 3e Offsets between pairs of replicate ^{14}C measurements on singe tree rings (error bars at 2σ; if there are more than two measurements each is successively plotted against the first). Figure 3f Offsets between pairs of replicate ^{14}C measurements on single-entity samples of charred plant remains (error bars at 2σ; if there are more than two measurements each is successively plotted against the first). Figure 3g Offsets between pairs of replicate ^{14}C measurements on calcined bone (error bars at 2σ).

Replicate Groups Containing Bulk Samples

During the 1990s the proportion of samples funded by English Heritage and dated by AMS increased from ca. 2% to ca. 75%. This meant that the proportion of bulk samples also reduced, as it was now usually possible to submit single entities for dating. Nonetheless, there were still some categories of material that had to be bulked even for dating by AMS, and conventional dating could still produce higher precision.

Eight replicate groups consisting of measurements on bulk samples of charred plant remains and measurements on single-entities from the same fired-feature are available. Three of these (38%) are statistically inconsistent at the 5% significance level, although overall ca. 88% of fragments appear to be freshly deposited in the context from which they were recovered.

Most bulk samples of short-lived charred plant material dated at this time thus probably produced results that are reasonably accurate.

Twenty-one replicate groups are available on bulk samples that were divided and dated more than once. The nine groups on charred plant remains dated by liquid scintillation spectrometry (LSS) are statistically consistent at the 5% significance level, although the experimental pair of samples from soot-blackened thatch is not. Five of the 12 groups on waterlogged plant macrofossils dated by AMS are also not statistically consistent at the 5% significance level (42%). This probably reflects inhomogeneity in the waterlogged plant remains recovered from sediment samples.

Replicate Groups on Sediment

Producing accurate chronologies for organic sediments is often challenging. Although there is frequently strong prior information in the form of relative sequence and depth information that can be incorporated in formal statistical modeling (Bronk Ramsey 2008; Haslett and Parnell 2008; Blaauw and Christen 2011), the taphonomic and chemical relationships between the organic material available for ^{14}C dating and the time of deposit formation are often complex.

The sediments considered in this study were almost all organic-rich deposits that formed in a variety of locations, from the inter-tidal zone to upland blanket bogs. Few lacustrine sequences or true soils were dated. There are a number of different fractions which may be selected for ^{14}C dating from such deposits:

- identifiable waterlogged plant macrofossils; thought to be from plants which grew on or around the sampled site as the sediment accumulated,
- acid insoluble, alkali soluble fraction of bulk sediment ("humic acid"): thought to derive from the decay of plant material that grew on the site as the sediment accumulated,
- acid and alkali insoluble fraction of bulk sediment ("humin"): thought to consist of the physical remains of the plant material that grew on the site, and
- solid fraction of bulk sediment that remains after the acid soluble fraction has been removed ("total organic" fraction): this consists of the "humic acid" and "humin" fractions combined.

The replicate dataset consists of 199 groups where the "humic acid" and "humin" fractions were dated separately (Figure 3b). Measurements on waterlogged plant macrofossils are also available from 36 of these groups.

Overall, 85 of the 199 replicate groups (43%) are statistically inconsistent at the 5% significance level. Similar reproducibility is apparent for both large samples processed for conventional dating (21 out of 57 replicate groups [37%] are statistically inconsistent at this significance level) and small samples dated by AMS (62 out of 140 replicate groups [44%] are statistically inconsistent at this significance level). In the cases where the replicate results on different fractions were inconsistent, the "humic acid" fraction generally returned younger ages (in 65 out of the 85 cases [77%]). Of the nine pairs which are offset by more than 1000 BP, five are of early Holocene age.

In 10 cases where there are results on waterlogged plant macrofossils as well as bulk fractions, the replicates are statistically consistent at the 5% significance level. In 12 cases results on

macrofossils join measurements on the bulk fractions that are statistically inconsistent at this significance level, and in 14 cases (39%) the results on the macrofossils are statistically inconsistent (at the 5% significance level) with statistically consistent pairs of measurements on the bulk fractions.

Variation in the homogenity of bulk samples of waterlogged plant macrofossils appears to be as great a risk in dating organic sediments as variations in the ^{14}C age of different chemical fractions of bulk sediment.

Replicate Groups on Carbonized Residues

Carbonized residues adhering to pottery sherds produced 56 groups of replicate measurements. All the samples were on the internal surfaces of the sherds and were interpreted as food remains from the use of vessels (although chemical characterisation of the dated material was not undertaken). Most sherds were considered to be close in age to the deposit from which they were recovered, either because they refitted with other sherds or because they were unabraded and fragile.

Of the 56 groups of replicate measurements, however, 36 (64%) produced results that are statistically inconsistent at the at the 5% significance level (Figure 3c). This type of sample is clearly problematic.

Four laboratories produced the measurements in these replicate groups, each using a different method of pretreatment. At the Oxford Radiocarbon Accelerator Unit (OxA-) generally the solid residue was dated following an acid-wash and multiple water washes (Brock et al. 2010); at Queen's University Belfast (UBA-) the solid residue was dated following an acid wash (Reimer et al. 2015); following an acid-base-acid pretreatment, the solid residue was selected for dating at the Scottish Universities Environmental Research Centre (SUERC-) (Dunbar et al. 2016) and the alkali-soluble fraction was selected for dating at Rijksuniversiteit Groningen (GrA) (Mook and Streurman 1983). The two pairs of results that are offset by more than 1000 BP include measurements pretreated by all these methods.

Each of these protocols produced measurements which are statistically consistent (at the 5% significance level) with measurements made using a different pretreatment. This is compatible with the conclusion of Hedges et al. (1992) that both acid insoluble/alkali soluble and alkali/acid insoluble fractions can provide accurate dates. But the dataset considered here suggests that around one in three of the measurements on carbonized residues is anomalous. This is a rather higher proportion than that suggested by Bayliss et al. (2011: fig 2.32), who suggest on the basis of the compatibility of dates on carbonized residues in Bayesian chronological models for Neolithic causewayed enclosures from southern Britain that around one in six such measurements are inaccurate.

That this is an issue arising from the chemical composition of some of the carbonized residues submitted for dating is suggested both by the fact that four different protocols each produce some measurements that are apparently correct and some that are clearly erroneous, and by the scale of the offsets observed between replicate measurements. These are not statistical outliers, but rather clearly samples from which all sources of exogenous carbon have not been successfully removed (Figure 3c).

Replicate Groups on Multi-Year Blocks of Wood

Replicate measurements are available on 91 samples of multi-year blocks of wood. Nineteen of these groups of results are statistically inconsistent at the 5% significance level (21%; Figure 3d). These replicate groups contain 198 measurements, and so at least 10% of reported values must lie beyond 2σ of the true value. This is double statistical expectation. One of these inconsistent groups consists of conventional measurements (1/5; 20%), two of a mix of conventional and AMS results (2/5; 40%), and 16 of AMS determinations only (16/81; 20%).

The replicate samples came from 29 archaeological sites, nine of which provided samples which produced inconsistent groups. Some sites, such as Star Carr, Yorkshire (Milner et al. 2018a) and Glastonbury Lake Village, Somerset (Marshall et al. forthcoming), produced a disproportionate number of problematic samples, probably because the waterlogged wood at these sites was particularly poorly preserved (Brunning 2013: 201–202; Milner et al. 2018b: 185–190). Samples from the Ferriby boats, Yorkshire had been conserved in glycerol in the 1940s before attempts at dating were undertaken in the 1990s (Wright et al. 2001). These issues with some of the dated samples probably explain the observed reproducibility.

Without these sites, six of the remaining 67 replicate groups (containing 120 results) are statistically inconsistent at the 5% confidence level (9%), which is in line with statistical expectation.

Replicate Groups on Single Tree-Rings

The replicate dataset consists of 89 groups where the same single tree-ring (as determined by dendrochronology) was dated separately. Of these replicate groups, which contain 192 results, 12 (14%) are statistically inconsistent at the at the 5% significance level (Figure 3e).

Many of the dated tree-rings fall within the period of single-year calibration data (Stuiver et al. 1998). Sufficient data are available to calculate the mean offset and associated standard error for three laboratories against these data: OxA-QL, 1.4 ± 7.9 BP (n = 25, average quoted error = 27 BP, samples dated 2007, 2012–2013, and 2016), SUERC-QL, 12.1 ± 8.0 BP (n = 26, average quoted error = 22 BP, samples dated 2007, 2012–13, and 2016), GrA-QL, −22.5 ± 13.8 BP (n = 6, average quoted error = 30 BP, samples dated 2007).

In each case, the mean offset is within two standard deviations of the Seattle data and within one standard deviation of the average quoted error on the measurements from that laboratory. Wiggle-matching of the series of measurements separately for each laboratory provides results that are compatible with the dendrochronological age of the tree-rings in the case of Jermyn Street (at 95% probability; Tyers et al. 2009: fig 6, the 2007 dataset) and Ledstone Hall (at 95% probability; Marshall et al. 2019: figs 3–4, the 2016 dataset). Single-year calibration data are available for this time range. Wiggle-matching the series of results from each laboratory separately (and together) provided consistent results that are very slightly younger than the ages known from dendrochronology for both Blanchland Abbey and Kilve Chantry (at 95% probability; Bayliss et al. 2017: table 5, the 2012–2013 dataset). This finding, however, appears to relate to the resolution of the available calibration data in this period rather than the accuracy of the reported measurements.

Replicate Groups on Single-Entity Samples of Charred Plant Remains

Sixty-six samples of single-entity charred plant remains have replicate measurements from more than one laboratory (Figure 3f). Of these groups, which contain 140 results, five (8%) are statistically inconsistent at the 5% significance level. This is in line with statistical expectation.

Replicate Groups on Single-Entity Samples of Waterlogged Plant Remains

Twelve replicate groups, including 25 measurements, are available on samples of non-charred plant remains. Two of these groups (17%) are statistically inconsistent at the 5% significance level.

Replicate Groups on Calcined Bone

Twenty-one samples of calcined bone have replicate measurements (Figure 3g), of which four are statistically inconsistent at the 5% significance level (19%). The true age of at least four of the 44 results must lie outside the reported 2σ range (9%), which is again in excess of the 5% expected on statistical grounds. All these samples were pretreated using acetic acid (Lanting et al. 2001) and consisted of a 2-g fragment of white calcined bone. The degree of calcination was only assessed on the basis of color.

Replicate Groups on Single-Entity Samples of Animal and Human Bone

No less than 400 samples of animal and human bone have replicate measurements from more than one laboratory (Figure 4).

Replication was undertaken in four cases to investigate whether PVA had been successfully removed during sample pretreatment, one replicate group was undertaken as part of investigating a known laboratory problem (a contaminated cylinder of oxygen at GU in 1993), and 10 samples were redated because the accuracy of legacy data was suspected. Overall, in 80% of cases the replicate groups are statistically inconsistent at the 5% significance level, confirming our suspicions about the original data.

The remaining 385 replicate groups, which contain 818 results, should represent a reasonably random sample of the animal and human bones dated for English Heritage/Historic England over the past 25 years. Overall, 70 groups of replicate results are statistically inconsistent at the 5% significance level (18%); six (out of 26) containing one or more conventional measurements (23%), and 64 (out of 359) containing only AMS measurements (18%).

Methods of bone pretreatment have varied both by laboratory and over the period covered by this dataset (cf. Whittle et al. 2011: tables 2.1–2.3), but basically can be divided into variants of that outlined by Longin (1971) and those than employ ultrafiltration (Brown et al. 1988). Of the 35 groups containing only AMS measurements prepared using variants of the Longin method, four are statistically inconsistent at the 5% significance level (11%); of the 115 groups containing AMS results obtained from samples prepared by both this method and with ultrafiltration, 26 are statistically inconsistent at the 5% significance level (23%); and of the 209 groups containing AMS results obtained from samples prepared by ultrafiltration, 34 are statistically inconsistent at the 5% significance level (16%). The greater variation in replicate groups including results produced by both approaches may reflect the much smaller proportion of intralaboratory replicates included in this category (5%), in comparision to the proportion for groups produced using variants of the Longin method

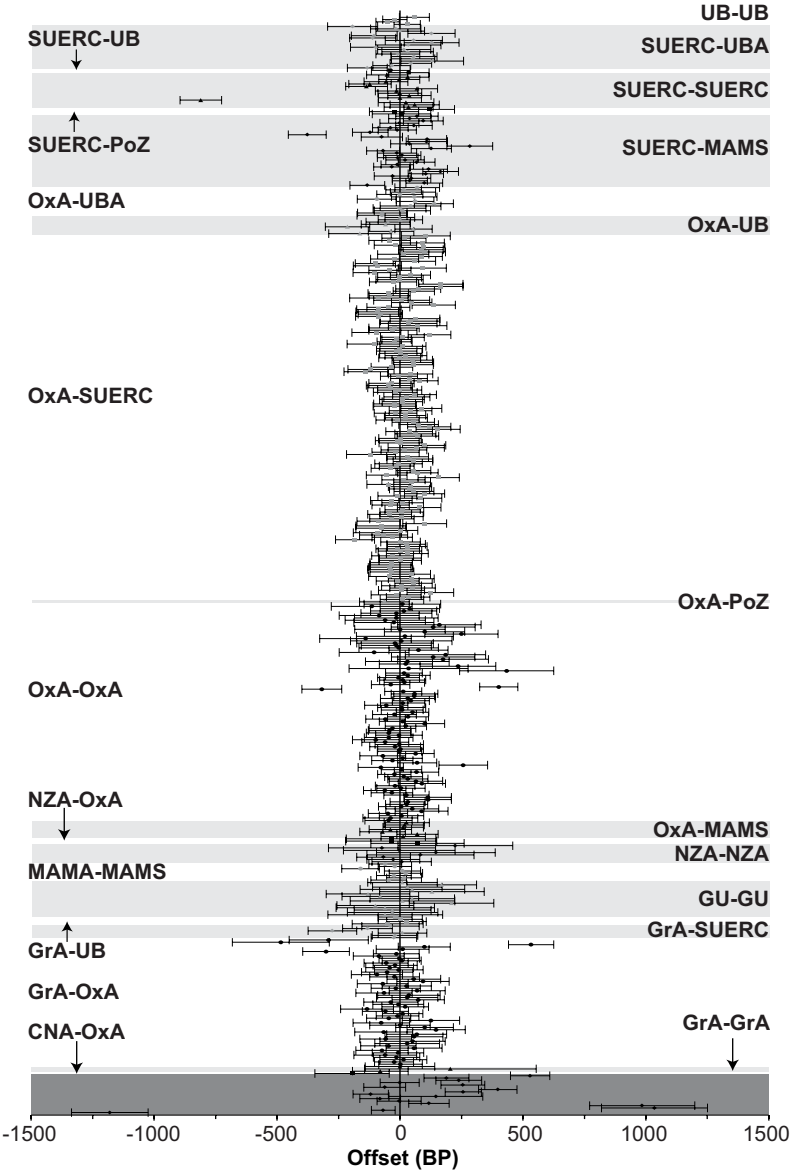

Figure 4 Offsets between pairs of replicate ^{14}C measurements on bone and antler (error bars at 2σ).

(77%) or ultrafiltration alone (27%). Of the six samples that produced replicate pairs of measurements that are offset by more than 500 BP, three had been consolidated with PVA and two tackled known laboratory problems.

Overall, the 359 replicate groups containing only AMS measurements include 765 results. At least 62 of these (8%) must lie more than 2σ from the reported value. This is slightly in excess of the 5% expected on statistical grounds.

Table 1 Risks in ^{14}C dating different archaeological sample types (from temperate climes and up to one-half-life in age). Archaeological risks have been assessed informally assuming best-practice in sample selection (cf. Bayliss et al. 2011: 56–58).

Sample material	Archaeological risk	Scientific risk
Pre-1993 measurements	Variable	12%
Sediments	Low	21%
Carbonized residues	Low	30%
Wood (multi-ring, mostly waterlogged)	Low	10%
Wood (single-ring, mostly from buildings)	Low	6%
Single-entity charred plants	High	4%
Waterlogged plants	Variable	8%
Bone and antler	Low	9%
Calcined bone	Low	9%

DISCUSSION

Accurate ^{14}C dating and accurate Bayesian chronological modeling are collaborative exercises between archaeologists and ^{14}C laboratories. In any dating program, there are risks on both sides. Table 1 summarizes these risks.

About one in 10 ^{14}C dates obtained before 1993 may be problematic, although bone samples dated at this time have a significantly higher chance of inaccuracy (ca. one in five). Difficulties with samples of carbonized plant material dated during these years probably relate more to the risks of bulk samples (Ashmore 1999) and age-at-death offsets in samples of unidentified charcoal than to problems in laboratory processing.

Overall a similar proportion (12%) of samples dated after 1993 fall outside the 2σ range, although there are clear variations by material type (Table 1). Accurate dating of sediment samples and charred residues on pottery sherds is clearly particularly challenging. About a fifth of replicate groups on the "humic acid" and "humin" fractions of bulk sediment are statistically inconsistent at the at the 5% significance level. In these inconsistent groups the "humic acid" fraction is generally considerably younger (mean difference -198 ± 23 BP), although in the statistically consistent groups there is no detectable bias (mean difference -9 ± 6 BP). This might suggest that the chronologies provided by statistically consistent replicate groups of measurements on different fractions of bulk sediment are robust, although almost half (48%) of waterlogged plant macrofossils dated from these consistent groups are not consistent with them at the 5% significance level. Dating sediments is clearly high risk. Misfits should be expected, and must be identified and managed. Replication is one part of the solution, but compatibility of the ^{14}C dates with the sequence and depths in the deposit are just as valuable. Again, ^{14}C scientists and their users need to work together. Accurate dating of carbonized residues on pottery sherds is clearly also problematic (Figure 3c), and in this case replication is often prevented by the limited material available. Again, archaeological information can be used to identify inaccurate results, but ultimately the solution may lay in alternative methods for directly dating the use of pottery (Casanova et al. 2018).

The other materials commonly dated by archaeologists are much less problematic. The replication of results from single rings from structural timbers in buildings and single-entity charred plant remains are within statistical expectation, although waterlogged wood and

Table 2 Mean difference and error for pretreatment methods between pairs of replicate ^{14}C measurements on bone.

Pretreatment	Statistically consistent at the 5% significance level (yr)	Statistically inconsistent at the 5% significance level (yr)
Longin	9 ± 11	234 ± 118
Ultrafiltration	6 ± 3	30 ± 23
Longin/ultrafiltration	−5 ± 5	−19 ± 44

waterlogged plant remains are slightly more challenging (Table 1). There are indications that this may be related to cellulose preservation on some sites, and we wonder whether laboratories need to develop procedures to characterize the preservation and yield of cellulose in wood samples analogous to those used by many for bone samples (e.g. van Klinken 1999).

If the observed reproducibility of ^{14}C results on waterlogged wood is poorer than we anticipated, that of bone samples is better. These prior expectations, as well as the character of the contexts from which the samples were obtained (many waterlogged wood samples come from structures in the inter-tidal zone, rather than the highly stratified sequences of articulating bone groups from excavations), are reflected in the number of samples sent for replication to two laboratories. Overall, ca. 8–9% of measurements on bone and antler samples probably lie more than 2σ from the true value, which is slightly in excess of statistical expectation. For this dataset, both pretreatment protocols derived from Longin (1971) and those employing ultrafiltration appear to be effective: no significant bias is apparent in the mean differences of either statistically consistent or statistically inconsistent groups (Table 2). Ultrafiltration may be more effective at reducing the scale of remnant contamination in a small number of problematic samples. It should be noted that this dataset consists of measurements on generally well-preserved bone from a temperate climate, which is predominantly less than one half-life in age. This reproducibility may not be obtained on older or poorly preserved material.

This study considers the reality of ^{14}C reproducibility for archaeologists over the previous research generation, when we have been working with measurements with an average quoted error of ± 36 BP. With the exception of some difficult material types, results are generally reproducible within this error. As counting errors reduce (Wacker et al. 2010), we must strive to maintain this accuracy as biases which have previously been hidden by larger counting statistics will become detectable and will need to be overcome.

ACKNOWLEDGMENTS

We would like to thank Mark van Strydonck, whose keynote lecture at the 8th International ^{14}C & Archaeology Symposium in Edinburgh in 2016 inspired us to identify as serial polygamists. Kate Cullen and Bisserka Gaydarska collated the data which enabled us to write this paper. Nothing would be possible without the persistent care and attention devoted on each and every sample by the technical staff at all the laboratories whose measurements are included in this study. We thank particularly the staff at Oxford and East Kilbride with whom we have had long-term collaborative relationships over the past quarter century.

REFERENCES

Ashmore PJ. 1999. Radiocarbon dating: avoiding errors by avoiding mixed samples. Antiquity 73(279):124–130.

Bayliss A. 2009. Rolling out revolution: using radiocarbon dating in archaeology. Radiocarbon 51(1):123–147.

Bayliss A, Bronk Ramsey C. 2004. Pragmatic Bayesians: a decade of integrating radiocarbon dates into chronological models. In: Buck CE, Millard A, editors. Tools for constructing chronologies: tools for crossing interdisciplinary boundaries. London: Springer-Verlag. p. 25–41.

Bayliss A, van der Plicht J, Bronk Ramsey C, McCormac FG, Healy F, Whittle A. 2011. Towards generational time-scales: the quantitative interpretation of archaeological chronologies. In: Whittle A, Healy F, Bayliss A, editors. Gathering time: dating the early Neolithic enclosures of southern Britain and Ireland. Oxford: Oxbow. p. 17–59.

Bayliss A, Beavan N, Hamilton D, Köhler K, Nyerges ÉÁ, Bronk Ramsey C, Dunbar E, Fecher M, Goslar T, Kromer B, Reimer P, Bánffy E, Marton T, Oross K, Osztás A, Zalai-Gaál I, Whittle A. 2016. Peopling the past: creating a site biography in the Hungarian Neolithic. Bericht der Römisch-Germanischen Kommission 94:23–91.

Bayliss A, Marshall P, Tyers C, Bronk Ramsey C, Cook G, Freeman SPHT, Griffiths S. 2017. Informing conservation: towards ^{14}C wiggle-matching of short tree-ring sequences from medieval buildings in England. Radiocarbon 59:985–1007.

Blaauw M, Christen JA. 2011. Flexible paleoclimate age-depth models using autoregressive gamma process. Bayesian Analysis 6(3):457–74.

Brock F, Ramsey CB, Higham TFG. 2007. Quality assurance of ultrafiltered bone dating. Radiocarbon 49(2):187–92.

Brock F, Higham T, Ditchfield P, Bronk Ramsey C. 2010. Current pretreatment methods for AMS radiocarbon dating at the Oxford Radiocarbon Accelerator Unit (ORAU). Radiocarbon 52(1): 103–112.

Bronk Ramsey C. 1994. Analysis of chronological information and radiocarbon calibration: the program OxCal. Archaeological Computing Newsletter 41:11–16.

Bronk Ramsey C. 2008. Deposition models for chronological records. Quaternary Science Reviews 27:42–60.

Bronk Ramsey C. 2009a. Bayesian analysis of radiocarbon dates. Radiocarbon 51(1):337–360.

Bronk Ramsey C. 2009b. Dealing with outliers and offsets in radiocarbon dating. Radiocarbon 51(3): 1023–1045.

Bronk Ramsey C, Higham TFG, Owen DC, Pike AWG, Hedges REM. 2002. Radiocarbon dates from the Oxford AMZ System: Archaeometry Datelist 31. Archaeometry 44(s1):1–150.

Brown TA, Nelson DE, Vogel JS, Southon JR. 1988. Improved collagen extraction by modified Longin method. Radiocarbon 30(2):171–177.

Brunning R. 2013. Somerset's peatland archaeology. Managing and investigating a fragile resource. Oxford: Oxbow Books.

Casanova E, Knowles TDJ, Williams C, Crump MP, Evershed RP. 2018. Practical considerations in high-precision compound-specific radiocarbon dating: eliminating the effects of solvent and sample cross-contamination on accuracy and precision. Analytical Chemistry 90: 11025–11032.

Dunbar E, Cook GT, Naysmith P, Tipney BG, Xu S. 2016. AMS 14 C dating at the Scottish Universities Environmental Research Centre (SUERC) Radiocarbon Dating Laboratory Radiocarbon 58(1):9–23.

Haslett J, Parnell A. 2008. A simple monotone process with application to radiocarbon-dated depth chronologies. Journal of the Royal Statistical Society: Series C (Applied Statistics) 57(4):399–418.

Hedges REM, Tiemei C, Housley RA. 1992. Results and methods in the radiocarbon dating of pottery. Radiocarbon 34(3):906–915.

Lanting JN, Aerts-Bijma AT, van der Plicht J. 2001. Dating of cremated bone. Radiocarbon 43(2A): 249–254.

Longin R. 1971. New method of collagen extraction for radiocarbon dating. Nature 230:241–242.

Mann WB. 1983. An international reference material for radiocarbon dating. Radiocarbon 25(2): 519–527.

Marshall P, Bayliss A, Farid S, Tyers C, Bronk Ramsey C, Cook G, Doğan T, Freeman SPHT, İlkmen E, Knowles T. 2019. ^{14}C wiggle-matching of short tree-ring sequences from post-medieval buildings in England. Methods in Physics Research B 438:218–226.

Marshall P, Brunning R, Coles J, Minnitt S, Bronk Ramsey C, Dunbar E, Reimer P. forthcoming. Stop me if you think you've heard this one before? The date of Glastonbury Lake Village, Somerset, UK. Antiquity.

McCormac G, Reimer P, Bayliss A, Thompson M, Beavan N, Brown D, Hoper S. 2011. Anglo Saxon Chronology Project. Laboratory and quality assurance procedures at the Queen's University, Belfast Radiocarbon Dating Laboratory. Portsmouth: English Heritage Research Department Report Series 89–2011.

Milner N, Conneller C, Taylor B. 2018a. Star Carr. Vol. 1. A persistent place in a changing world. York: White Rose University Press.

Milner N, Conneller C, Taylor B. 2018b. Star Carr. Vol. 2. Studies in technology, subsistence and environment. York: White Rose University Press.

Mook WG, Streurman HJ. 1983. Physical and chemical aspects of radiocarbon dating. In: Mook WG, Waterbolk HT, editors. Proceedings of the First International Symposium ^{14}C and Archaeology: PACT 8. p. 31–55.

Naysmith P, Scott EM, Cook GT, Heinemeier J, van der Plicht J, van Strydonck M, Bronk Ramsey C, Grootes PM, Freeman SPHT. 2007. A cremated bone intercomparison study. Radiocarbon 49(2): 403–408.

Reimer PJ, Hoper S, McDonald J, Reimer R, Svyatko S, Thompson M. 2015. The Queen's University, Belfast: laboratory protocols used for AMS radiocarbon dating at the ^{14}CHRONO Centre. Portsmouth: English Heritage Research Report 5-2015.

Scott EM. 2003. The third international radiocarbon intercomparison (TIRI) and the fourth international radiocarbon intercomparison (FIRI) 1990–2002: results, analyses, and conclusions. Radiocarbon 45(2):135–408.

Scott EM, Naysmith P, Cook GT. 2017. Should archaeologists care about ^{14}C intercomparisons? Why? A summary report on SIRI. Radiocarbon 59(5):1589–1596.

Staff RA, Reynard L, Brock F, Ramsey CB. 2014. Wood pretreatment protocols and measurement of tree-ring standards at the Oxford Radiocarbon Accelerator Unit (ORAU). Radiocarbon 56(2): 709–715.

Stuiver M, Reimer PJ, Braziunas TF. 1998. High-precision radiocarbon age calibration for terrestrial and marine samples. Radiocarbon 40(3): 1127–1151.

Tyers C, Sidell J, van der Plicht J, Marshall P, Cook G, Bronk Ramsey C, Bayliss A. 2009. Wiggle-matching using known-age pine from Jermyn Street, London. Radiocarbon 51(2): 385–396.

van der Plicht J, Wijma S, Aerts AT, Pertuisot MH, Meijer HAJ. 2000. Status report: The Groningen AMS facility. Nuclear Instruments and Methods in Physics Research B 172(1):58–65.

van Klinken GJ. 1999. Bone collagen quality indicators for palaeodietary and radiocarbon measurements. Journal of Archaeological Science 26:687–695.

Wacker L, Bonani G, Friedrich M, Hajdas I, Kromer B, Němec M, Ruff M, Suter M, Synal HA, Vockenhuber C. 2010. MICADAS: routine and high-precision radiocarbon dating. Radiocarbon 52(2):252–262.

Ward GK, Wilson SR. 1978. Procedures for comparing and combining radiocarbon age determinations: a critique. Archaeometry 20(1):19–32.

Whittle A, Healy F, Bayliss A. 2011. Gathering time: dating the early Neolithic enclosures of southern Britain and Ireland. Oxford: Oxbow.

Willis EH, Tauber H, Münnich KO. 1960. Variations in the atmospheric radiocarbon concentration over the past 1300 years. Radiocarbon 2:1–4.

Wright EV, Hedges REM, Bayliss A, Van de Noort R. 2001. New AMS radiocarbon dates for the North Ferriby boats—a contribution to dating prehistoric seafaring in northwestern Europe. Antiquity 75(290):726–734.