[RADIOCARBON, VOL 32, NO. 3, 1990, P 341-346]

A CONSIDERATION OF SOME BASIC IDEAS FOR QUALITY ASSURANCE IN RADIOCARBON DATING

ROY SWITSUR

Cambridge University, Godwin Laboratory, Free School Lane, Cambridge CB2 3RS, UK

ABSTRACT. Most radiocarbon ages are readily accepted by researchers in all disciplines. It is recognized, however, that discrepancies appear in the literature. These problems have been highlighted by the International Collaborative Study. The introduction of quality control and assurance techniques used in some laboratories for many years could reduce or eliminate aberrant results. I present here some of the basic considerations of this approach in the processes of conventional radiocarbon dating.

Amongst the results of the International Collaborative Program (Scott *et al* 1988), it was disturbing to discover that the ages for one of the test samples, reported by investigators from 18 radiocarbon laboratories, ranged from 490 ± 60 BP to 1670 ± 70 BP, *ie*, a spread of ca 1180 years. Further, the distribution of these ages appeared quite regular throughout the range and showed no obvious satisfactory grouping from which to derive the 'correct' or consensus age. A simple mean or even a weighted average would be inappropriate. It is difficult to ascribe to this range and distribution any obvious statistical justification. The extremes of the age spread represent a difference in relative radiocarbon content of almost 13%. Now most, if not all, of the workers should be capable of measuring relative radiocarbon content to better than 1% and, always assuming that the test sample was homogeneous, it is necessary to attempt to account for the discrepancies in the findings. Since the assays involved were made for a special test sample, we may presume that they were carried out with at least as much care as for normal samples requiring age determinations. The result is obviously disturbing in view of the many diverse studies that rely on radiocarbon dating.

Although disconcerting, these divergent results are not so surprising inasmuch that a scan through the radiocarbon literature reveals similar discrepancies of various magnitudes amongst samples that have been multiply dated. For example, workers in six reputable dating laboratories performed 18 age determinations on wood from Chelford, Cheshire (Worsley 1980). The finite ages reported ranged between 26,200 and 60,800 BP, a spread of 34,600 years; also, some were given as infinite. At the more recent part of the time scale, scientists in two different laboratories found consistent age differences of over 300 years between samples from the same Roman/late Iron Age body of 'Lindow Man'. Similarly, a group of samples for a Neolithic trackway in the Somerset Levels, England, gave ages which, although internally consistent, were much earlier than expectations based on palynologic and stratigraphic evidence and the ages of similar trackways in the region. It is interesting to ponder on the number of other such anomalies that might be found if multiple age determinations were made at more sites. Often, an unexplained outlier, which can be perhaps a millennium earlier, is found in an otherwise close group of determinations. Disparities of smaller numerical values also appear with some regularity. Probably 10 to 20% of ^{14}C ages do not agree with archaeologic or geologic expectations, but this rate of agreement is no worse than the results of other accepted dating methods, eg, thermoluminescence, electron spin resonance and potassium/argon dating. The reasons for these obvious discrepancies are generally not explained, and the earliest age is usually accepted on the grounds that later ones may be due to incomplete removal of more recent carbon contamination. This explanation is probably not defensible in the case of outliers, where misassociation of the sample is a possibility. Despite these aberrations, most researchers generally accept that the overall quality of ^{14}C age determinations is comparable with that of other dating methods. Although checks are often difficult to make, this

judgement is based on independent criteria. Radiocarbon scientists at different laboratories sometimes collaborate on a single site, a reasonable strategy that should serve to detect possible bias. An example of this is the good agreement on split samples in tests on Neolithic monuments on Orkney (Switsur & Harkness 1979). It is, of course, important that we make continuous efforts to maintain high standards in the quality of ^{14}C dates. Both scintillation spectrometry and gas proportional counting can produce excellent and comparable results. Pearson and Stuiver (1986) amply demonstrated this by their work on the ^{14}C time-scale calibration.

A closer consideration of the results of the intercomparison study shows that some participants often were able to reproduce their determinations and readily detect duplicates. That is, both the precision and the bias for each worker were consistent. These observations hold the clue to the distribution of results produced. It seems probable that a major part of problem may lie in the values for the standards and backgrounds used in the age calculations. Other things being equal, too high a value for the background would produce too great an age, whereas too low an activity value for the contemporary oxalic acid standard would give an age that was too young. The activity ratio used in the age calculation can be very sensitive, especially in the case of smaller sized samples, to relatively small changes in the values of the constants. Any contamination introduced during processing into a radioactively dead sample would certainly increase the measured background. Similarly, organic contamination, apart from nuclear bomb carbon, introduced during the preparation of the contemporary oxalic acid would reduce its activity (assuming that a correction for $\delta^{13}C$ is made, otherwise such isotopic fractionation could introduce an age error of up to 4% in younger samples (Nehmi 1980)). Scientists who work with both gas proportional and liquid scintillation spectrometry tend to use repeatedly a given oxalic acid preparation. A contaminated contemporary standard automatically introduces bias into the results of the age determination. This is consistent with the findings of the international study and can explain both the age spread and the internal reproducibility observed. Frequent preparation of fresh standards and background samples combined with the use of statistical control graphs (Switsur 1990) to check the operation of the system would give warnings and help avert many of these problems.

It is a self-evident presumption that all uncontaminated specimens must have some fixed *actual* or true age. However, the measurements performed in a radiocarbon age determination are essentially those of experimental physics. Consequently, because these measurements inevitably involve some type of error of observation, the *exact* age of the specimen will be, to that extent, indeterminate. The practical problem is to reduce these errors to as low a level as is compatible with equipment stability, resolution and specimen integrity. The experimental or observational errors arise from different causes and follow no simple laws. We find, quite generally in physics and other 'exact' sciences, that repeated measurements by the same scientist using the same equipment on a given specimen does not always produce the same result. This may be due to lack of uniformity in the performance of the equipment or to the variability on the part of the user or possibly to small changes in other factors that control the measurements. The errors, then, may be systematic or accidental, and may be divided conveniently into two categories - those that are mainly statistical and hence may vary randomly in occurrence, magnitude and in sign, and those that produce bias, which may also possibly occur at random, arising from the equipment, the standards, the specimen, or they may be observer-dependent. All these are troublesome; much time and effort is spent in attempting to discover and eliminate the causes. Nevertheless, it is not reasonable to imply a guarantee of the success of these efforts. The closer the results of the determination to the actual age of the specimen the greater is the *accuracy* of that determination. Except in the instance of known-age test samples, the accuracy will be unknown. Repetition of the measurements will, in general, tend to produce a statistical spread of results around the 'most

probable' age. The magnitude of this spread is a measure of the *precision* of the result. High precision and accuracy can ensue from careful evaluation of the component measurements required for the determination. Important amongst these are instrumental background and contemporary radiocarbon standard, as well as the radioactivity of the specimen itself. The stable carbon isotopic ratio measurement of the specimen is sufficiently precise that it inconsequentially affects the overall precision of the sample age. Other factors can also affect the precision - the fine tuning and the stability of the electronic system, the effect of varying barometric pressure on the counting, the purity of the counting medium, the efficiency of the counting medium, and, in liquid scintillation counting, the selection of the counting vials, the composition of the scintillation cocktail, matching the scintillation light output to the photomultiplier photocathode sensitivity, and so on. Careful monitoring of, or control over, these numerous interdependent procedures and processes is essential for reliable and reproducible results.

The concept of quality control or assurance is by no means new in physical and chemical determinations and it has been exercised for many years by some members of the radiocarbon community. However, quality control has rarely extended beyond statistical processes. Statistical quality control may be a more appropriate term, leaving quality assurance to the non-statistical aspects. In this sense, it is possible to check that the observable parameters are within statistically reasonable limits and to attempt to return them to within these limits should some large deviation occur. We can also adjust the system to adapt to any new, changed values (eg, an abrupt change in counter background). Radiocarbon activity is a good example of where these statistical controls may be properly applied. Earlier researchers were well aware of the shortcomings of their apparatus when they attempted some of the most sensitive radiometric measurements ever made in physics. Only by careful statistical controls was it possible to produce results of reasonable reliability. Environmental shielding was often inadequate, and even a small fluctuation of laboratory temperature affected electronic stability. It was very difficult to sustain reliable counting over a sufficiently long period with statistically respectable results. That it was accomplished was to no small extent due to careful quality control procedures. With the introduction of improved apparatus, the task of radiocarbon determination has lightened. We now expect that scientists in laboratories with the latest, very stable electronic equipment and specially shielded counters should have little difficulty in surpassing results of only a decade ago.

In radiocarbon measurements of the natural environment, problems are rather different from those confronting radiochemists dealing with 'tracer' levels of radioactive substances, which may have activities higher than environmental samples by 2 or 3 orders of magnitude. Typical tracer studies involve tracking a radioactive substance through various stages of a process, to detect or identify its presence, rather than the more demanding *measurement* of concentration, as is the case in age determinations. Precise counting parameters are then, less critical in tracer experiments. At the extreme range of conventional radiocarbon dating, an activity as low as 0.006 picocuries per gram of carbon needs to be measured to reach 57,000 years (10 half-lives). It is obviously desirable to work with as high a signal-to-noise ratio as possible as long as this is compatible with high stability and high efficiency. Some of the latest techniques suggested for greatly enhancing the $S\sqrt{B}$ ratio of scintillation spectrometers are at the expense of efficiency and have tended to reduce stability. Thus, this type of development may not be appropriate for dating; low *stable* background and *stable* counting parameters should be the goal.

Of the various types of error indicated earlier, only statistical errors may be evaluated quantitatively. Researchers often produce graph representations of time series of counter background measurements or other standards to demonstrate visually the stability of their systems. These can be deceptive. It is more useful to construct *statistical control graphs* (Switsur 1990) which enable us to examine more thoroughly the behavior of the system. However, this procedure

may require more counting of reference samples, which could involve laboratory scheduling in terms of both time and budget. The detection of non-random or perhaps seasonally periodic changes in the standards requires these to be counted on a reasonably regular, though not necessarily frequent, basis. Statistical quality control graphs are based on the powerful technique of the analysis of variance. 'Rational' data subgroups are taken and two calendric graphs are plotted, one of the subgroup means and one of the mean range within the subgroup. Checks are made for any significant difference between the means of the subgroups. A value related to the standard deviation is also obtained from the subgroups, rather than from the square root of the total number of counts, as is customary in radiocarbon dating. The latter leads to a vanishingly small error term after a large number of measurements have been pooled, which is not realistic. Statistical control graphs are equally valuable for the quantitative display of the variability of the standard NBS oxalic acid or background. This graph technique provides an excellent monitor for the variability of the counting system.

From the viewpoint of the physics, as Libby propounded in his theory of radiocarbon dating, the values representing the extremes of the radioactivity scale, *ie*, accurate and precise knowledge of the activity of the contemporary standard and the background, are all that are necessary and sufficient to define the dating system. In practice, however, the standards may not be sufficiently well known and we prefer to make additional checks for bias by other means. A variety of substandards are used. Some workers perform comparative assays on either historically known-age or well-dated samples such as wood from Caligula's boat from Lake Nemi or multiply checked Allerød material. Others prefer known-age samples prepared from dendrochronologically dated wood. Often check samples are chosen for proximity to more sensitive regions, such as the contemporary standard, eg, ANU sucrose, or background. As an additional check on stable operation, some chronologists using liquid scintillation monitor a sealed radiocarbon standard some 10 or 20 times the modern activity in relation to a known-age sample. Statistical control graphs prepared from the measurements of these ratios would provide a sensitive check on instrumental stability and aid removal of bias. Nevertheless, these *ad hoc* check samples are not very widely used and the amounts available and the range covered are relatively small: universally available samples are needed.

The results of the International Collaborative Study have highlighted the problems, but international intercomparisons do not provide complete answers to comparability amongst the researchers, though, without doubt, some of the participants have profited by improving their procedures. However, more than three years have elapsed since the inception of the latest study and the publication of the results. This is far too long to be of any real assistance to most research scientists, who need to be able to correct laboratory errors rapidly after their detection. What is urgently required is a readily available suite of known-activity samples covering a wide age range, enabling the chronologist to draw on them at will as reference samples to check, as rapidly as possible, the full operational procedures of his/her laboratory.

Samples of 'known age' in the form of radioactive benzene became available some years ago from an intercomparison study by British radiocarbon scientists (Otlet *et al* 1980). They ranged from twice the contemporary activity to an apparent 20,000 years and good agreement was obtained among the participating groups. These demonstrated the usefulness of the idea of a range of available samples of known activity in the diagnosis of possible faults in the counting system. Despite the utility of these known activity samples, benzene would be difficult to distribute widely. With the right kind of samples, however, both radioactivity measurements and sample preparation techniques could be checked, difficulties identified and eliminated. This sort of procedure is important if the flow of dating samples is not to be impeded.

The AQCS program, organized by the IAEA in Vienna, already provides authenticated

reference materials of stable isotopes for use in this manner and physicists using them publish their results when they believe they have eliminated their laboratory problems. Participating groups have thus established a consensus of data for each specimen. The IAEA is now planning to establish a very similar scheme for the radiocarbon community, once the difficulty of producing appropriate homogenized and authenticated specimens has been overcome. This project is timely, very welcome and should go a long way to help solve some of our problems. Nevertheless, since equipment, procedures and personnel in laboratories change, problems will regularly occur and continuous use of quality control graphs and checks with known-age samples should be an essential agenda for every dating research laboratory. This apparently fruitless effort will inevitably be reflected in the output of the laboratory since each measuring system (counter or spectrometer) must be included and it is essential that proper provision must be made in the budget. The result should be more accurate and precise radiocarbon ages.

Statistical control graphs and known-age check samples do not cover the complete processes in radiocarbon dating, for the work of the research scientist involves projects with real samples of varied nature and sedimentary origin where quality assurance is equally vital, but which aspect is often overlooked. High-precision radiocarbon analysis requires high-precision samples. The majority of radiocarbon determinations are carried out as integral parts of university interdepartmental research projects, primarily in the geosciences. The radiocarbon physicist is usually a member of an interdisciplinary research team. The type and extent of involvement of the co-workers are diverse and determined at the outset according to each member's specialty. Each is unlikely to have expertise in the others' fields and the project is designed around their individual experience. From the radiocarbon viewpoint, the unique information that the radiocarbon determinations will provide in helping solve the problem needs to be evaluated. Quality assurance must extend to the field and specifically to the identification and acquisition of specimens. Sites should be carefully chosen and samples selected for authenticity and compatibility with the aims of the project. When possible, the radiocarbon scientist should aid in securing critical samples as is the practice for TL, ESR or K/A determinations, for his judgement of suitable sample materials and identification of possible contamination will be the keenest. In archaeologic projects, which account for ca 10 to 15% of radiocarbon determinations, the tendency is to employ commercial dating laboratories for sample assays, possibly because few radiocarbon laboratories are found in archaeology departments. In this case, the responsibility for checking sample suitability and recording details for possible contaminants lies with the collector. Unfortunately, archaeologists are usually not trained in these techniques and frequently submit samples without adequate documentation of sample material, matrix or provenience. Consequent inappropriate pretreatment and/or incomplete removal of contaminants may lead to a false age.

Sample contamination is potentially a source of large errors. It is important that sample preparation should be subject to quality assurance and performed in a laboratory clean room, preferably with a positive pressure and dust-free air circulation to avoid the introduction of modern contamination. The equipment should be dedicated to radiocarbon research, not general departmental use. Unfortunately, many radiocarbon laboratories do not have these elementary provisions. Similarly, many facilities are not well equipped for comprehensive sample testing and pretreatment. Even moderately complex procedures involving the extraction of pure substances for dating may be beyond their equipment capability. From the results of the International Comparison Study, some of the discrepancies indeed seem to stem from inadequate preparation or purification. For example, some participants reported great difficulty in removing chloride from a humic-acid sample. It is important that all dating laboratories have access to ancillary purification equipment and quality reagents. Without quality assurance in this area, all the time-consuming and expensive statistical controls and bias checks are in vain.

In a properly designed project, the outcome of the radiocarbon determinations are of critical importance to the research and not merely expensive extra data. The radiochronologist is responsible for decisions on materials or fractions that would produce the most reliable ages and, in the final publication, is closely involved with the interpretation of the results he has obtained. It is essential that quality assurance should work, both in the field, in proper sample acquisition and verification, as well as in the laboratory, in good specimen preparation and measurement technique. The International Collaborative Study has thrown into high relief the fallacies of attempts at cut-price dating with inadequate apparatus and facilities. We now urgently need to invest in university radiocarbon laboratories, to provide high-quality equipment and well-trained personnel, in order to collaborate in worthwhile research programs.

REFERENCES

Nehmi, VA 1980 Isotopic fractionation of NBS oxalic ^{14}C standard and its effects on calculated age of materials. *In* Stuiver, M and Kra, RS, eds, Internatl ^{14}C conf, 10th, Proc. *Radiocarbon* 22(2): 501-504.

Otlet, RL, Walker, AJ, Hewson, AD and Burleigh, RM 1980 ^{14}C interlaboratory comparison in the U K: Experiment design, preparation and preliminary results. *In* Stuiver, M and Kra, RS, eds, Internatl ^{14}C conf, 10th, Proc. *Radiocarbon* 22(3): 936-946.

Pearson, GW and Stuiver, M 1986 High-precision calibration of the radiocarbon time scale, 500-2500 BC. *In* Stuiver, M and Kra, RS, eds, Internatl ^{14}C conf, 12th, Proc. *Radiocarbon* 28(2B): 839-862.

Scott, EM, Aitchison, TC, Harkness, DD, Baxter, MS and Cook, GT 1988 An interim progress report on stages 1 and 2 of the International Collaborative Program. *In* Long, A and Kra, RS, eds, Internatl ^{14}C conf, 13th, Proc. *Radiocarbon* 31(3): 414-421.

Switsur, VR 1990 Statistical quality control graphs in radiocarbon dating. *Radiocarbon*, this issue.

Switsur, VR and Harkness, DD 1979 The radiocarbon dates. *In* Renfrew, AC, ed, *Investigations in Orkney*. London, Thames & Hudson, Soc Antiquaries: 70-73.

Worsley, P 1980 Problems of the radiocarbon dating of the Chelford Interstadial in England. *In* Cullingford RA, Davidson, DA and Lewin, J, eds, *Timescales in geomorphology*. New York, John Wiley & Sons: 289-304.