Nuclear Instruments and Methods in Physics Research B 525 (2022) 62–66

Contents lists available at ScienceDirect

Nuclear Inst. and Methods in Physics Research, B

journal homepage: www.elsevier.com/locate/nimb

What lies behind radiocarbon intercomparisons and the design of the new intercomparison, GIRI?

E.M. Scott [a,*], P. Naysmith [b], G. Cook [b]

[a] School of Mathematics and Statistics, University of Glasgow, Glasgow G12 8QW, United Kingdom
[b] SUERC, Radiocarbon Dating Laboratory, East Kilbride, United Kingdom

ARTICLE INFO

Keywords:
Intercomparison
Quality assurance
Reference materials

ABSTRACT

Given the complexity of the radiocarbon dating process, the diversity of materials being dated, the continued technical developments, GIRI (the Glasgow international radiocarbon intercomparison) is the next development of the series of intercomparisons to support continuing quality assurance. GIRI has been designed to continue this programme and to meet a number of objectives, including the most fundamental one, to provide an independent assessment of the analytical quality of the laboratory/measurement and an opportunity for a laboratory to participate and improve (if needed). The principles that we followed in the creation of GIRI are to provide. A) A series of unrelated individual samples, spanning the dating age range B) Some samples linked to earlier intercomparisons to allow traceability C) Some known age samples, to allow independent accuracy checks D) A small number of duplicates, to allow independent estimation of laboratory uncertainty E) Two categories of samples, bulk and individual to support laboratory investigation of variability. All of the GIRI samples are natural (wood, peat and grain), some are known age, and overall their age spans approx. >40,000 BP to modern. Ultimately, we wish to define consensus values for all the samples and a quantified uncertainty supporting a more in-depth evaluation of laboratory performance and variability.

1. Introduction

Radiocarbon dating is one of the most widely used dating techniques in archaeology and geochronology being used to provide estimates of the ages of artefacts, when and for how long archaeological sites were occupied and the timing and nature of environmental change. It is a remarkable tool, in that, since its discovery in the 1940s it has become the cornerstone of much archaeological and environmental research, with a corresponding growth in the numbers of laboratories set up to provide measurements. Like every complex measurement process, each measurement has an uncertainty, sometimes described as its error, which fundamentally is a quantification of the variability that would be observed were we able to make true repeated measurements. Contributions to this uncertainty come from measurement of standards (of known activity), backgrounds (no ^{14}C activity), known activity reference materials (which allows the evaluation of bias which is sometimes included in the definition of uncertainty (described as being systematic) and other sources, which potentially include technician effects, pretreatment effects as well as other contributions not easily uniquely identified. Given the complexity of the processes, the diversity of materials being dated, and ongoing technical developments, there has been a sustained effort based in part on a series of intercomparisons to fully quantify the uncertainties on the reported age, accounting for all laboratory processes. Such intercomparisons form an important part of a quality assurance framework common in many other areas of science.

2. Metrological concepts

Given the challenges of the dating process, starting from the archaeological context and field sampling, and then the complexities of the measurement process, quality assurance and quality control processes are critical and intertwined with the concept of measurement accuracy and precision and uncertainty quantification. Some of the key metrological concepts include bias, accuracy and precision, repeatability and reproducibility. These concepts are critical in ensuring a well calibrated measurement system and part of that also comes from benchmarking of measurements since it is highly likely that scientific studies will require comparability of results from different laboratories. As a result of the needs to deliver accurate and precise measurements but also as part of general, good laboratory practice, including laboratory

* Corresponding author.
E-mail address: Marian.scott@glasgow.ac.uk (E.M. Scott).

https://doi.org/10.1016/j.nimb.2022.06.015
Received 19 April 2022; Received in revised form 21 June 2022; Accepted 23 June 2022
Available online 5 July 2022

benchmarking and quality assurance, the ^{14}C community has undertaken a wide-scale, far reaching and evolving programme of global intercomparisons, to the benefit of laboratories and users alike [1]. Each intercomparison has been designed to meet a number of objectives, including the most fundamental one, to provide an independent assessment of the analytical quality of the laboratory/measurement and an opportunity for a laboratory to participate and improve (if needed). Comparability of measurements made by different laboratories or in the same laboratory but at different times has been critical in the IntCal programmes, culminating in the most recent IntCal20 revision [2]. Here measurements have been made in different trees, speleothems and other carbon repositories, in different laboratories using different equipment and have then been brought together to revise the global calibration curves. Indeed in the development of the new curve fitting methodology [3], results from several previous intercomparisons were used to inform specification of key modelling parameters but the authors also noted *"Even when the same sample is measured in different laboratories, we have evidence of a greater level of observation spread (i.e. over-dispersion) in the 14C measurements within objects from the same calendar year than the laboratory reported uncertainties would support (see* e.g. *Scott et al. 2017). Recognizing this potential overdispersion, whatever its cause, is important both for curve estimation and resultant calibration for IntCal20*".

At the core of any intercomparison lies the samples, and a further important objective of radiocarbon intercomparisons is the creation of a set of recognised reference materials which are well characterised. Reference materials can be used regularly and allow a laboratory to fully explore its own processes and procedures.

2.1. Reference materials and primary standards

For many years, the importance of reference materials has been recognised, as in this quote from 1977. "Our primary concern is with the use of reference materials and reference methods in transferring accuracy and precision throughout large multi-laboratory measurement networks. Although reference materials and reference methods can be effectively used to assure accuracy and precision within an individual laboratory, the great demand for compatible national measurement systems has led to their increased utilization in a total systems approach to accurate measurements in a variety of areas of chemical analysis".[4].

Within the metrology literature, a distinction is drawn between primary and secondary reference materials, the critical point being that secondary reference materials may be used on a daily basis, and considered as working standards. A primary reference material is one "that is designated or widely acknowledged as having the highest metrological qualities and whose value is accepted without reference to other standards of the same quantity, within a specified context". A secondary reference material or standard is one "whose value is assigned by comparison with a primary standard of the same quantity" [5].

Within the radiocarbon community, laboratories have often created their own working standards e.g. humics, cellulose or barley mash e.g. [6]. One advantage of a *secondary* reference material created as part of an intercomparison is that its activity has been verified in many laboratories.

2.2. Uncertainty, precision and error- a hierarchical quantification of variation

The basic measurement model assumes that a ^{14}C age determination X_i is subject to variability due to some known and potentially unknown laboratory and environmental sources, and that this is quantified in the so-called counting error, which is an uncertainty quantification unique to the individual age determination. It is perhaps simplest to consider the counting error as the culmination or propagation of a series of contributing uncertainties or sources of variation- forming a hierarchy. Focussing on the AMS context, measurements made in the same batch (AMS wheel) may share common error components (e.g. same standard or backgrounds being used within an AMS target wheel, same operator, same operating mode). An important practical and philosophical point is, if the determination X_i were repeated many times (say n), with all conditions apparently being kept absolutely constant, then there would be a distribution of results $X_i \sim N(\mu, \sigma^2)$, typically assumed a Normal distribution. It would then be possible to estimate the dispersion, σ^2 in the distribution of the set of measurements by the sample variance, providing an empirical quantification of the measurement precision. The interesting and relevant question is how the sample variance in such a set relates to the typical quoted error values. As well of course, we are not able to make absolute replicate measurements since we cannot control all sources of variability within the sample or the laboratory or the machine. Ref. [7] provide a detailed investigation of uncertainty within their laboratory using a combination of different experiments.

2.3. Replication and repeatability

If a laboratory were able to make multiple measurements on the same sample, we would observe a distribution of ages, and if we quantified the standard deviation of the distribution then we would obtain a measure of the repeatability of a measurement. This is where the value of the secondary reference material becomes clear- a comparison of the routinely quoted individual error with the repeatability error assessed as described above using the reference material allows an experimental assessment of whether there is any over dispersion (and whether the quoted individual error should be adjusted). A commonly used statistical measurement model, allows any age determination to have an additional and independent source of potential variability beyond that being reported, either in an additive model, or in a multiplicative model (hence the widely used error multipliers [8] and sometimes described as the external uncertainty.

A widely used statistic in this setting is the reduced χ^2 statistic [9] or mean square weighted deviation (MSWD) which allows a formal evaluation of the existence of over-dispersion.

2.4. Accuracy and offsets

When the age of a material is known, then the measured age can be compared to the known age (of course the known age is likely to be in calendar years, while the measured age will be in radiocarbon years (years BP) so needs to be calibrated). Accuracy is a statistical concept, which pragmatically means that on average the measured values give the correct age, within uncertainty. The difference between the correct age and the mean measured value is the offset. Typically, every laboratory will include known age material to be routinely measured, and again the potential value of the reference material which has been well characterised is clear, any laboratory offset can be experimentally determined.

3. Laboratory intercomparisons- the principles and practice

Laboratory intercomparisons (sometimes called proficiency trials or round-robins) are internationally recognised and often officially organised as linked to laboratory accreditation schemes. The decision to undertake a focussed intercomparison is based often on a number of reasons, including when setting up a new laboratory, or when investigating the effects of different pre-treatment protocols. Global intercomparisons that aim to recruit widely from the laboratory community often have different objectives, such as establishing standards and reference materials (with known activities/concentrations). The principles of the design and stages in a proficiency trial are detailed and discussed briefly below. We use the current (GIRI) intercomparison as the exemplar.

3.1. Sample materials

For the design of an intercomparison, either natural or synthetic samples can be used [10]. There are challenges in both types of material. For synthetic materials, they may not represent the materials which are typically dated and may be difficult to distribute, however they can be formed in sufficient quantity to meet all needs, and their ^{14}C activity can be pre-defined. For natural materials, they are selected as representative of routinely dated materials, but they can be challenging to acquire in sufficient quantity and their ^{14}C activity is not always known in advance. In the vast majority of cases, the Glasgow organised intercomparisons have used samples which are routinely dated materials, including wood (often known age decadal tree rings and more recently single tree rings), bone, peat, and shell. The main criteria for selecting samples are that they should, (1) span the spectrum of age (modern to background) and material, (2) satisfy rigorous homogeneity testing, and (3) be known age where possible.

With the increasing interest in single tree-ring radiocarbon measurement, and the resolution this offers into rapid events within the global (or local) carbon cycle, tree-rings continue to play an important part in our intercomparisons.

A significant practical challenge comes from the need to provide sufficient material for 80+ laboratories and ideally to have sufficient material remaining for future use. These considerations mean that material must be sourced in bulk, which raises concerns about sample homogeneity especially when routine measurements are now made on a few milligrams of material.

3.2. Which stage of the dating process?

Since the radiocarbon dating process can be considered in a hierarchical, additive manner, we can also consider which stage of the dating process the intercomparison sample relates to, since that indicates what component of uncertainty/variability the sample addresses. Most samples require the laboratory to pre-treat the sample using their own procedures, before conversion to CO_2 and graphite target- this would be the full accounting of uncertainty (excluding the environmental sampling). Some samples however may already have been pretreated, and as such they address the conversion/graphite phase of the dating process. Samples have in common either that they must be converted to graphite (by the individual laboratory method) or gas (for the gas ion source) and then measured. The advantages of such a hierarchical design lie in the potential to evaluate the different components of variation and to quantify them [11].

3.3. Past global intercomparison samples

In Scott et al (2018) [1] an overview of more than 30 years of global intercomparisons was provided. Wood has been one of the most frequently used materials, including known age (dendro-dated) samples, and even if not dendro-dated, then it is still possible to provide identical tree-ring series thus avoiding issues about homogeneity and comparability of material provided to each laboratory. Examples have included dendro dated tree rings from the master chronologies of Belfast and Germany, and archaeological samples. Such samples have all required pre-treatment, but in addition, cellulose has also been provided avoiding issues of pre-treatment effects. Peat has been a commonly used material since it is possible to sample in bulk, but it can be challenging to demonstrate homogeneity unless there are known age horizons which limit the sampling frame. Pretreatment to the humic acid fraction is an important process overcoming homogeneity issues, and humic acid has been a commonly distributed material. Bone is a less commonly used material, partly due to the complexities of pre-treatment methods. Anchoring the modern end of the radiocarbon timescale, we have made use of barley mash, with a known year of growth. Background and near background samples are also important. While many laboratories have process blanks (eg graphite), past intercomparison have sourced and included natural background and near background samples, such as Kauri wood samples and background bone.

3.4. Selected examples of some specific intercomparisons

We have chosen 6 examples of intercomparisons to illustrate their function and diversity. In the first case, an intercomparison was used as part of the commissioning of a new laboratory to allow an evaluation of performance [12]. The second and third examples focus on a specific material, namely bone [13,14], while a fourth study looks at re-measuring of original materials in an evaluation of historical dates made in the late 1940s [14]. The fifth and sixth studies [16] and [9] relate to the IntCal exercises (with a focus on tree rings).

Turney et al. [12] presented results from a small intercomparison to investigate the accuracy of the new Chronos C-14 facility. A set of contiguous tree-ring cellulose samples across selected time periods were measured in Chronos and 3 other facilities. This intercomparison is an example built round the evaluation of a new facility and its comparability to other laboratories.

Jull et al. [15], performed a new series of measurements on samples that were part of early measurements made in 1948–1949, and included the measurement of several samples in 4 different laboratories. Their results showed good agreement to the early results despite the different technologies. This study is particularly interesting since in the 60 years since those early measurements were made there have been very significant technological changes in the dating process.

Naysmith et al. [13], designed and planned an intercomparison that studied charred bones, investigating the differences due to the different laboratory protocols and pretreatment procedures.

Huels et al. [14], undertook an intercomparison study of a bone close to background, focussing on the comparison of individually prepared and measured bone collagen radiocarbon activities.

These two intercomparisons focussed on a very specific material, one which is less commonly dated and for which there are a number of pretreatment challenges.

Wacker et al. [9], reported a single tree-ring intercomparison "to systematically test the reproducibility of AMS measurements on wood samples following discussions at the IntCal meeting in Belfast in 2016". Three sets of consecutive single tree-ring samples from different time intervals were used and 16 laboratories participated.

Manning et al. [16] investigated the comparability of low-level gas proportional counting to AMS measurements, in the context of the Northern Hemisphere IntCal curve, where there were repeat measurements. They concluded that interlaboratory variation is relevant and potentially a dominant issue.

In conclusion, these 6 examples illustrate some of the different features of intercomparisons, the resulting discoveries and the fuller appreciation and quantification of the uncertainties associated with radiocarbon dating.

4. GIRI samples and design

The most recent intercomparison GIRI was delayed from 2019, samples were dispatched on 2021 with results expected in 2022. More than 70 laboratories received samples, the vast majority being AMS facilities. Building on our previous work and experiences, all of the GIRI samples are natural (wood, peat and grain), some are known age, and overall their age spans approx. >40,000 BP to modern. In the case of peat, the sample has been pre-treated to humic acid, and we also include a cellulose sample, but other samples require pre-treatment. The complete list of sample materials includes: humic acid, whalebone, grain, a number of single ring dendro samples, a number of dendro-dated wood samples spanning a number of rings (e.g. 10 rings), background and near background samples of bone and wood.

The principles that we followed in the creation of the study design

are

A) A series of unrelated individual samples, spanning the dating age range
B) Some linked samples to earlier intercomparisons to allow traceability
C) Some known age samples, to allow independent accuracy checks
D) A small number of duplicates, to allow independent estimation of laboratory uncertainty
E) Two categories of samples, *bulk* and *individual* to support laboratory investigation of variability

5. GIRI design principles

Referring to the two groups of samples (*bulk* and *individual*), *individual* group samples are typical of the samples provided in previous intercomparisons, where the volume of material is sufficient to make at most a very small number of repeat measure. The second group, the *bulk* samples provide a quantity of material, sufficient to allow AMS labs to run (and report) multiple measurements from different wheels/batches over the space of a six months experimental phase. It is intended that sufficient material will remain to allow labs to use these as internal quality assurance samples. There are two such *bulk* samples provided. The first (wood) will typically require pretreatment, while the second (humic) will require little or no pretreatment.

5.1. Why this design?

The purpose of including the *individual* samples is to allow each laboratory to quality check (once consensus values and uncertainties have been defined), their laboratory operation at the time of analyses (so a classical proficiency trial). The group of *bulk* samples provides laboratories with well characterised materials which can function as reference materials, in sufficient quantity to be run routinely and thus allow assessment of both laboratory precision and accuracy.

5.2. GIRI analysis plan

Analysis of the GIRI results will have several strands: reporting on individual laboratory performance, and sample characterisation.

5.2.1. Sample consensus values

In the first instance, we will define consensus values for all the samples with a quantified uncertainty. We will follow the procedures described in [1] to provide the sample consensus value and its associated uncertainty.

In addition, based on the GIRI design, we will also be able to quantify:

5.2.2. Laboratory disparity

On basis of the duplicate samples. This quantity is simply the unsigned difference of the duplicates divided by the square root of the sum of squares of the quoted errors. Values for the disparity measure of >1 indicate that the discrepancy between duplicate samples was greater than expected given the quoted errors. The disparity data may be used by the laboratory to assess their analytical reproducibility.

5.2.3. Laboratory offset

Is an estimate of the difference between the lab result and “true value” for a specific sample. More typically, the true is a robust measure (median) of the consensus on all the study results. Laboratory offset data allows us to assess the extent of interlaboratory variation and the existence of systematic biases. Offsets can be calculated relative to the consensus value, or where we have an independent age estimate, relative to that quantity. This latter calculation will typically be based on dendro-dated wood samples [17].

5.2.4. Excess variation

Traditionally, evaluation of z-scores, is a standard approach to evaluate the performance relative to the consensus value [1], but of particular interest in this context is the variability in the results and checking of the measurement uncertainties. Here we use a zeta score and evaluate the chi-squared statistic χ^2 (which is the sum of the squared zeta-scores). The zeta score is interpreted similar to the z-score but includes the uncertainty on the consensus value.

It is also common to evaluate a reduced χ^2 (sometimes also called the Mean Weighted Squared Deviations). The reduced χ^2 is the χ^2 divided by n–1 (where n is the number of observations used in the calculation of the consensus value). We compare the reduced χ^2 value to 1, values greater than 1 would indicate over dispersion in the results around the consensus value.

5.2.5. Traceability

Where we have used samples that have previously been evaluated, then we will incorporate the new measurements and update/revise as required any consensus values.

6. Conclusions and discussion

This short paper has presented an overview of the current GIRI intercomparison and placed it in the context of the 30 year history of intercomparisons and also highlighted its differences to more recent studies. It has also reflected on the benefits of creating new reference materials from the intercomparison programme and very briefly considered some of their challenges. Such archived reference materials offer rich resources for new laboratories and for commissioning new instruments. An archive of material has been formed and is available to the community on request from the authors. Benefits to participating laboratories have been identified including benchmarking, identification of systematic offsets and additional sources of variation.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

References

[1] E.M. Scott, P. Naysmith, G.T. Cook, Why do we need 14Cintercomparisons?: The Glasgow 14C intercomparison series, a reflection over 30 years, Quat. Geochronol. 43 (2018) 72–82.
[2] P.J. Reimer, et al., The IntCal20 Northern Hemisphere radiocarbon age calibration curve (0–55 kcal BP), Radiocarbon 62 (4) (2020) 725–757.
[3] T.J. Heaton, M. Blaauw, P.G. Blackwell, C. Bronk Ramsey, P.J. Reimer, E.M. Scott, The IntCal20 approach to radiocarbon calibration curve construction: a new methodology using Bayesian splines and errors-in-variables, Radiocarbon 62 (4) (2020) 821–863.
[4] G.A. Uriano, C.C. Gravatt, G.H. Morrison, The role of reference materials and reference methods in chemical analysis, Crit. Rev. Anal. Chem. 6 (4) (1977) 361–412.
[5] ISO. Guide 30: 1992/Amd 1:2008, terms and definitions used in connection with reference materials. 2nd ed. Geneva: ISO; 2008.
[6] P. Naysmith, E.M. Scott, E. Dunbar, G.T. Cook, Humics - their history in the radiocarbon intercomparisons studies, Radiocarbon 61 (5) (2019) 1413–1422.
[7] A.T. Aerts-Bijma, D. Paul, M.W. Dee, S.W.L. Palstra, H.A.J. Meijer, Meijer H A J (2021) An independent assessment of uncertainty for radiocarbon analysis with the new generation high-yield accelerator mass spectrometers, Radiocarbon 63 (1) (2021) 1–22.
[8] E.M. Scott, G.T. Cook, P. Naysmith, Error and Uncertainty in Radiocarbon Measurements, Radiocarbon 49 (2) (2007) 427–440.
[9] L. Wacker, E.M. Scott, A. Bayliss, D. Brown, E. Bard, S. Bollhalder, M. Friedrich, M. Capano, A. Cherkinsky, D. Chivall, B.J. Culleton, M.W. Dee, R. Friedrich, G.W. L. Hodgins, A. Hogg, D.J. Kennett, T.D.J. Knowles, M. Kuitems, T.E. Lange, F. Miyake, M.-J. Nadeau, T. Nakamura, J.P. Naysmith, J. Olsen, T. Omori, F. Petchey, B. Philippsen, C. Bronk Ramsey, G.V.R. Prasad, M. Seiler, J. Southon, R. Staff, T. Tuna, Findings from an in-depth annual tree-ring radiocarbon intercomparison, Radiocarbon 62 (4) (2020) 873–882.

[10] E.M. Scott, (ed), The third international radiocarbon intercomparison (TIRI) and the fourth international radiocarbon intercomparison (FIRI) 1990-2002: results, analyses, and conclusions, Radiocarbon 45 (2003) 135–408.
[11] E.M. Scott, T.C. Aitchison, D.D. Harkness, G.T. Cook, M.S. Baxter, An overview of all three stages of the international radiocarbon intercomparison, Radiocarbon 32 (3) (1990) 309–319.
[12] C. Turney, L. Becerra-Valdivia, A. Sookdeo, Z.A. Thomas, J. Palmer, H.A. Haines, H. Cadd, L. Wacker, A. Baker, M.S. Andersen, G. Jacobsen, K. Meredith, K. Chinu, S. Bollhader, C. Mario, Radiocarbon protocols and first intercomparison results from the CHRONOS 14-carbon-cycle facility, university of New South Wales, Sydney, Australia, Radiocarbon 63 (3) (2021) 1003–1023.
[13] P. Naysmith, E.M. Scott, G.T. Cook, J. Heinemeier, J. van der Plicht, M. Van Strydonck, C.B. Ramsey, P.M. Grootes, S.P.H.T. Freeman, A cremated bone intercomparison study, Radiocarbon 49 (2) (2007) 403–408.
[14] M. Huels, J. van der Plicht, F. Brock, S. Matzerath, D. Chivall, Laboratory intercomparison of pleistocene bone radiocarbon dating protocols, Radiocarbon 59 (5) (2017) 1543–1552.
[15] A.J.T. Jull, C.L. Pearson, R.E. Taylor, J.R. Southon, G.M. Santos, C.P. Kohl, I. Hajdas, M. Molnar, C. Baisan, T.E. Lange, R. Cruz, R. Janovics, I. Major, Radiocarbon dating and intercomparison of some early historical radiocarbon samples, Radiocarbon 60 (2) (2018) 535–548.
[16] S.W. Manning, B. Kromer, M. Cremaschi, M.W. Dee, R. Friedrich, C. Griggs, C. S. Hadden, Haddens C S Mediterranean radiocarbon offsets and calendar dates for prehistory, Science Advances 6 (12) (2020).
[17] E.M. Scott, G.T. Cook, P. Naysmith, R.A. Staff, Learning from the wood samples in ICS, TIRI, FIRI, VIRI and SIRI, *Radiocarbon* 61 (5) (2019) 1293–1304.