

SECTION 10: SUMMARY AND CONCLUSIONS

A significant and substantial effort has been made by the ^{14}C community in quality assurance (QA) procedures, of which participation in FIRI is only one part but one that provides an independent and blind check on laboratory performance. The overwhelming willingness to participate is a testament to the importance which laboratories place on quality.

FIRI provides a spot-check of operational performance at the time it was carried out. It is not intended to be a means to create a league chart of laboratories. FIRI does not measure consistent performance over a period of time and this is one reason why the FIRI results are published without attribution to laboratories. Feedback is provided to laboratories, which they may choose to act upon, and it has been the case that a number of laboratories have identified and corrected problems as a result of participation in FIRI.

Derived reference values for the FIRI materials have been obtained and are given in Table 10.1.

Table 10.1 Consensus values and estimated standard error

Sample	Known age	Consensus value (estimated 1 σ precision)
AB (pMC)	—	0.24 pMC[a] (95% CI [0.23 – 0.30])
C (yr BP)	—	18,176 (10.5) yr BP[b]
DF (yr BP)	3200–3239 BC (^{14}C age 4495 BP)	4508 (3) yr BP
E (yr BP)	—	11,780 (7) yr BP
GJ (pMC)	—	110.7 (0.04) pMC
H (yr BP)	313–294 BC (^{14}C age 2215 BP)	2232 (5) yr BP
I (yr BP)	3299–3257 BC (^{14}C age 4471 BP)	4485 (5) yr BP

[a]pMC = percent modern carbon
[b]^{14}C years before present (yr BP) is 1950

The findings of FIRI are best considered in terms of some of the overall design aims for FIRI. These can be expressed in the form of some general questions posed and answered in the following.

10.1 HOW COMPARABLE ARE THE LABORATORIES?

Comparability can be considered in terms of the average result and also in terms of the variation in results.

On Average

We find overall, and on average, no evidence of significant differences between AMS, GPC, and LSC laboratories, with the exception of the near-background Kauri sample, where, on average, the age reported by AMS laboratories is highly likely to be older.

Variation

In terms of the variation reported, we find that more LSC laboratories reported results identified as extreme or outliers. Outliers were also identified for GPC laboratories, but in less number. In addition, there is some evidence as a group that the overall variation in the results is less for AMS laboratories; however, there are several factors which may, in part, explain this result, namely: a) feeder labs may have used a common AMS facility, and b) given the sample size requirements,

https://doi.org/10.1017/S0033822200032665 Published online by Cambridge University Press

AMS facilities are able to prepare multiple targets and quote average results, which would be expected to show less variation.

10.2 HOW VARIED ARE THE RESULTS AND WHAT FACTORS EXPLAIN THE VARIATION?

Components of Variation

The components of variation can be considered in two ways: random and systematic.

Random

Random components of variation would be apparent from the amount of scatter in the results, which shows no specific pattern but perhaps manifests itself through either outliers or anomalous values. The chances of outliers occurring, assuming they occur by chance, is roughly 1 in 20. For an individual laboratory, random variation might also manifest itself in that the difference between duplicates or the difference between the measurement and the known age tend to be larger than would be expected, given the laboratory quoted error.

Thus, in 10 results (as in FIRI), the presence of 1 outlier in the set is not unlikely and, therefore, does not indicate a problem. However, more than 1 outlier in a set of 10 is increasingly unlikely, assuming that such observations occur by chance.

Systematic

Systematic components of variation are apparent as a shift or offset in the results, i.e., results are always too high or always too low relative to a known age (or a reference value). Possible reasons could be incorrect estimation of the background, calibration of the modern standard, or a source (constant) of modern or very old carbon within the laboratory.

In the analysis of the FIRI results, we see evidence of both random and systematic sources of variation.

- Roughly 10% of the total results are identified as outliers (which is around twice as frequent as would be expected). Yet, it should be noted that the distribution of outliers is not uniform across the laboratories, with the majority of outliers coming from around only 14% of the laboratories. The distribution of outliers across samples is uniform, so no one sample material is more varied than any other.
- Comparing laboratory results to both dendro-dated samples and the derived reference values for the materials, we find evidence for small laboratory offsets relative to the derived reference values for some laboratories.

10.3 CAN WE IDENTIFY ANY REASONS FOR THE VARIATION IN RESULTS?

We have studied the effect of the modern standard and background material used and found no evidence that these factors make a significant contribution to the overall variation observed. For some samples, we did see evidence of an effect of pretreatment (FIRI C). Issues of outlier identification showed that they were often associated with the modern standard used by the laboratories and this is a recurring theme in much of the analysis.

https://doi.org/10.1017/S0033822200032665 Published online by Cambridge University Press

10.4 ACCURACY AND PRECISION ISSUES

Precision

Within the measurement process, the quoted error is a measure of precision on the measurement. Ideally, it quantifies the variation to be expected in the measurement were it to be repeated many times. For the radiometric laboratories, its basis is the Poisson counting/decay process, although other sources of random variation are also typically included in its calculation. For the AMS laboratories, the quoted error is not based on the Poisson decay process, but its interpretation remains similar to that for the radiometric laboratories.

The duplicate samples included in FIRI allow estimation of precision (without issues of true age being considered).

From the 3 sets of duplicate results, we see that, on average, the difference in duplicates is zero (for all laboratories and also for individual laboratories), but the magnitude of difference is frequently large relative to the quoted errors (and larger than expected given the interpretation of the quoted error). The implication is that a source of variation is not completely accounted for in the quoted error in these cases. In a number of cases, we also see evidence of agreement between the duplicates, which is, in fact, better than would be expected given the quoted errors.

Accuracy

Accuracy is concerned with the "correctness of the result." Ideally, with exactly known-age samples, this could be independently estimated (for our dendro-dated samples, the true ^{14}C age is not known exactly, but only within a range, due to that fact that it is measured). The master measurements are based on decadal samples, which do not correspond exactly to the samples provided in FIRI. This range could be as much as 100 yr, which corresponds to twice a commonly-quoted error value.

For our materials, we must assume that we can define (through calculation) what the "true" ^{14}C age will be (the consensus value), and then, we can estimate for each laboratory whether there is a constant offset from this consensus (hence, a measure of accuracy).

This is not an ideal situation since the issue of precision of the estimate of the consensus value should also be considered. However, the consensus value is based on a large number of results and so its precision is high, relative to the individual measurements.

We found evidence that a number of laboratories had small, but significant, offsets relative to the consensus profile. One possible explanation is that of mis-estimation of the background or modern standard activity, but other reasons are possible. Results from FIRI do not allow further examination of this.

Overall, the evidence supports the fact that ^{14}C laboratories are generally accurate and precise, but that notwithstanding internal QA procedures, some problems still occur that can best be detected by participation in intercomparisons such as FIRI. The results from FIRI are significant in that they show a broad measure of agreement between measurements made in different laboratories on a wide range of materials. They also demonstrate no statistically significant difference between measurements made by radiometric or AMS techniques.

Finally, some of the same features identified in FIRI were also observed in the previous exercise ("Part 2" of this issue). This reinforces the idea that an extra, independent check on laboratory per-

https://doi.org/10.1017/S0033822200032665 Published online by Cambridge University Press

formance is required, and suggests that internal QA procedures, while essential, do not address all QA issues. When advised of the analysis, laboratories are able to instigate a number of corrective measures and we would anticipate that FIRI would result in similar activity.

There is a clearly demonstrated need for standards and reference materials to which laboratories have ready access to allow checking and correction. As a result of FIRI (and previous Glasgow-led programs, especially TIRI, see Part 2), a small archive of natural materials has been created for use by the ^{14}C community. Some of the materials are extremely limited and sufficient remains exist for AMS measurement only, while for some others, a substantial store exists. However, information concerning its existence has been, and is being, disseminated to laboratories, with the purpose that such samples could be used to check laboratory performance.

Table 10.2 Archived material from TIRI

Sample identifier	Description	Activity range
H	Ellanmore peat	Less than 3 half-lives
A	Barley mash	Modern
J	Buiston Crannog wood	Less than one half-life
G	Fuglaness wood	More than 5 half-lives
I	Travertine	—
K	Turbidite	More than 3 half-lives
F	Doublespar	Background
L	Whalebone	More than 2 half-lives

Table 10.3 Archived material from FIRI

Sample identifier	Description	Activity range
A,B	Kauri wood	More than 5 half-lives
C	Turbidite	More than 3 half-lives
E	Humic acid	Less than 3 half-lives
G, J	Barley mash	Modern
H	Dendro-dated wood	Less than 1 half-life
I	Dendro-dated cellulose	Less than 1 half-life
K	Dendro-dated cellulose	Before bomb

10.5 FURTHER INTERCOMPARISONS

At the end of FIRI, a small follow-up questionnaire was circulated to all ^{14}C laboratories, seeking their views on the intercomparison just completed and any comments they might have on future requirements and organization. Of those who responded, around 80% thought the FIRI workload had been sufficient, that the timescale was sufficient, that there was sufficient sample material, and that the feedback had been timely and in sufficient detail.

The view on the frequency of intercomparisons was roughly split, with 50% thinking that an interval of 4–5 yr was optimal and 37.5% preferring an interval of 3 yr.

A total of 33% thought that there should be fewer than 10 samples, while 59% thought 10 samples was reasonable.

With respect to the anonymity of the laboratories, 44% thought laboratories should be anonymous, 41% thought it should be up to the individual laboratory, and 16% thought laboratories should not be anonymous.

An overwhelming majority said they would participate in a future intercomparison (94%).

From the experience gathered from both TIRI and FIRI and also the response to the questionnaire, it would appear that the ^{14}C community is fully supportive of intercomparisons and see them as benefiting them greatly. Thus, further intercomparisons are seen as an essential part of the community's QA.

Design Issues in Future Intercomparisons

There are a number of design issues relating to the organization of a laboratory intercomparison. Many relate to the sample material, but also there are issues concerning the conduct of the trial, which are briefly discussed below.

Sample Material

There are 2 options in the selection of material: first, all samples are of a single class of material (e.g., only shell or peat or wood), this of course limits the ability to generalize the results, so more commonly for ^{14}C dating at least, the materials used have been representative of routinely-dated material.

Activity Range

The activity or age of the test samples should cover the applied ^{14}C timescale.

Sample Size and Homogeneity

A key question, especially when using natural samples, is the homogeneity of the material, which should be tested. Obviously, as sample requirements in terms of weight may vary quite widely (through differences in pretreatment procedure, counting, and technique), it is necessary that the sample should be demonstrably homogeneous at the finest level required. This is an important issue as there is ever growing requests for dates from smaller and smaller samples.

Number of Samples

The number of samples is balanced between the needs of the statistical analysis of the data and, of course, the practical commitments of the participating laboratories. Preferably, numbers of test samples should be greater than 4 and there should be replication (with the identity of duplicate pairs withheld from the participating laboratories). The presence of duplicate samples allows a direct assessment of a laboratory's repeatability or the within-lab variation.

Perceived Needs

All of the previous ^{14}C intercomparisons have provided valuable information to laboratories, and hence, to users. As a result, it is clear that such checks as FIRI and others are, and will continue to be, necessary and that they must operate in addition to any within-laboratory procedures. Nor is it clear in these previous studies that the increased availability of an extensive range of reference materials has presented an immediate solution to the problem of laboratory comparability, as might have been hoped. Increasing the scope of reference materials and standards is important, since by their inclusion, the dating determinations can be better constrained, but only if laboratories make regular use of them in routine operation. There is increasing pressure to date smaller (even to the molecular level) and older samples. More conventional laboratories are forming close collaborations with accelerator laboratories, which has meant developing in-house techniques for

target preparation. Thus, an accelerator laboratory may have a number of target preparation laboratories providing it with targets presenting new issues of comparability. However, perhaps the most significant factor is that as we strive to measure smaller and smaller samples, the issue of sample homogeneity becomes more and more important; indeed, the definition of a sample becomes critical. In some of the studies already completed in which AMS laboratories have participated, some evidence of sample in-homogeneity has been reported, which the conventional laboratories were not able to detect. There are difficulties in taking a representative sub-sample from the bulk of material; indeed, how do we know it is representative? We do not fully know the potential scale of natural ^{14}C variation in sample matrices.

10.6 FUTURE PROPOSALS

Continuation in this work is important. The linkage to previous work provides an invaluable continuity (e.g., IAEA and other reference materials are still available and should be used), but further, new materials should be sought, including known-age material, and certainly a "background" organic sample is essential. For the conventional laboratory, the typical sample requirement might be 5gC, with sample age ranges from 1 to 4 half-lives. However, for the AMS laboratories and for those conventional laboratories where small samples are dated, we need to explore the natural variation in reportedly single-event samples (deposits of charcoal, grain from a single growing season, single insects from a well-defined stratum). This information is not just important for the laboratory, but is also of fundamental importance for the sample submitter who must select samples referring to the event of interest. There are new challenges for ^{14}C dating in continuing to ensure the quality of results.

Discussion with laboratories and results from the questionnaire responses have indicated a general desire for further intercomparisons of this more classical nature such as FIRI. Yet, in consultation, an additional proposal has been drawn up such that there should be a rolling 4-yr program. A major intercomparison such as FIRI would be organized every 4 yr, but in each of the 3 preceding years, a small number of samples (e.g., 3) would be sent to laboratories to be analyzed in a short time and feedback given within a short follow-up period. In this way, the "spot-check" nature of FIRI and the lack of continuous monitoring of performance would be remedied. This would be of more use to the participating laboratories, but would also provide a better guarantee of quality assurance to the user communities. Plans are currently being drawn-up for the next intercomparison (VIRI) and will be presented at the 2003 Radiocarbon International Conference in New Zealand.

REFERENCES FOR FIRI, SECTIONS 1–10

Bland JM, Altman DG. 1986. Statistical methods for assessing agreement between two methods of clinical measurement. *Lancet* (i):307–10.

BS5497. The organisation of collaborative trials. British Standards Institute.

Gulliksen S, Scott EM. 1995. TIRI report. *Radiocarbon* 37(2):820–1.

Hogg AG, Higham T, Robertson S, Beukens R, Kankainen T, McCormac FG, van der Plicht J, Stuiver M. 1995. Radiocarbon age assessment of a new, near background IAEA ^{14}C quality assurance material. *Radiocarbon* 37(2):797–805.

ISG. 1982. An interlaboratory comparison of radiocarbon measurements in tree rings. *Nature* 198:619–23.

Long A, Kalin RM. 1990. A suggested quality assurance protocol for radiocarbon dating laboratories. *Radiocarbon* 32(3):329–34.

Mangerud J, Svendsen JI, Astakhov VI. 1999. Age and extent of the Barents and Kara ice sheets in Northern Russia. BOREAS 28(1):46–80.

Polach H. 1989. ^{14}Care. *Radiocarbon* 31(3):422.

Rozanski K, Stichler W, Gonfiantini R, Scott EM, Beukens RP, Kromer B, van der Plicht J. 1992. The IAEA ^{14}C intercomparison exercise 1990. *Radiocarbon* 34(3):506–19.

Scott EM, Aitchison TC, Harkness DD, Cook GT, Baxter MS. 1990. An overview of all three stages of the international radiocarbon intercomparison. *Radiocarbon* 32(3):309–19.

Scott EM, Harkness DD, Miller BF, Cook GT, Baxter MS. 1992. Announcement of a further international intercomparion exercise. *Radiocarbon* 34(3):528–32.

Scott EM, Harkness DD, Cook GT. 1998. Interlaboratory comparisons: lessons learned. *Radiocarbon* 40(1): 331–43.

Sementsov AA, Zaitseva GI, Gorsdorf J, Nagler A, Parzinger H, Bokovenko NA, Chugunov KV, Lebedeva LM. 1998. Chronology of the burial finds from Scythian monuments in southern Siberia and Central Asia. *Radiocarbon* 40(1):713–21.

Wilson SR, Ward GK. 1981. Evaluation and clustering of radiocarbon age determinations: procedures and paradigms. *Archaeometry* 23(1):19–39.