Commercial Assays for Serum Osteocalcin Give Clinically Discordant Results

Paul W. Masters, Richard G. Jones, Donald A. Purves, Edward H. Cooper, and John M. Cooney

Serum samples from 9 healthy controls and from subjects with primary hyperparathyroidism (n = 5), Paget disease (n = 3), pregnancy (n = 5), glucocorticoid therapy (n = 5), postmenopausal osteoporosis (n = 10), and renal failure (n = 10) were used to assess the clinical agreement among eight commercially available assay kits for osteocalcin (OC). These kits differ in their assay configurations (six radioimmunoassays, two immunoradiometric assays), standards (five bovine, three human), and antibodies (six polyclonal, two monoclonal). Individual results were divided by the mean OC of the control subjects for each assay and expressed as percentage deviations. The expected wide variation in absolute OC concentrations between kits was only partially reduced by this transformation. Agreement was equally poor when absolute OC concentrations were compared with the reference ranges quoted by the manufacturers. The discordance was particularly marked in renal failure, presumably because of immunoreactive fragments, and in osteoporosis. Systematic differences could not be attributed to assay format, species source of standard, or antibody specificity. We conclude that results cannot be compared between assays even when normalized against healthy subjects, and that standardization is needed.

Indexing Terms: variation, source of/standardization/intermethod comparison/osteoporosis/renal failure/bone/proteins

Osteocalcin (OC), or bone Gla protein, is the most abundant noncollagenous protein in mature human bone, where it constitutes 1–2% of the total protein. A 49-amino-acid protein of 5800 Da, it is synthesized by osteoblasts and incorporated into bone matrix, although a small proportion is released into the circulation, where it may be measured by immunoassay (1). Since its discovery in the mid-1970s (2), OC has become the focus of much interest, because its concentration in serum appears to be a specific index of bone synthesis (3). Serum OC has been reported as useful in assessing both progression and response to treatment in postmenopausal osteoporosis (3–6), in predicting the histological type of renal osteodystrophy (7), and in assessing glucocorticoid-induced bone suppression (8–11).

Until recently, most studies have used in-house assays for OC. This, together with the lack of an international standard preparation, has hindered both the interpretation of results and the widespread use of OC measurements. A multicenter study by Delmas et al. (12), which also included two commercial assays, concluded that absolute OC concentrations cannot be directly compared between laboratories, even when the same standard preparation is used for calibration.

Since that assay standardization report was published, several more commercial assays have become available. This development, in conjunction with the favorable clinical studies, will put pressure on laboratories to include OC in their repertoires. However, the differences in assay format, antibodies, and standards among manufacturers suggest that results will be no more comparable than between different in-house assays (most of which are based on bovine standards and rabbit antisera), and probably will be less so. In the present study, we assessed the use of these assay kits in a variety of clinical states, using methods similar to those of Delmas et al., who compared assays in terms of multiples of the mean value for healthy controls (12).

Materials and Methods

Clinical Groups

The procedures followed were in accordance with the standards of the Ethical Committee of the Leeds General Infirmary. Blood was collected from 9 healthy adult volunteers (8 men, 1 woman, ages 28 to 65 years) and from 39 patients with clinical states previously reported as being associated with high or low serum OC. These were:

1. Primary hyperparathyroidism (PHPT). Five patients with hypercalcemia (albumin-adjusted plasma calcium >2.5 mmol/L, reference range 2.13–2.43) and inappropriately raised intact parathyroid hormone (PTH) (assayed with Allegro, Nichols Institute Diagnostics, Geneva, Switzerland) with normal plasma creatinine concentrations.

2. Paget disease of bone. Three patients with Paget disease diagnosed on clinical and radiological grounds with plasma alkaline phosphatase more than three times the upper limit of the adult reference range. Blood was also collected from a fourth patient, who had received bisphosphonate therapy.
4. Glucocorticoid therapy. Five adult patients taking prednisolone 5–40 mg/day for at least 3 days for either rheumatoid arthritis (three) or asthma (two).
5. Postmenopausal osteoporosis. Ten women (mean age 65 years) referred to an osteoporosis clinic with normal plasma calcium and without increased plasma creatinine or alkaline phosphatase.

Sample Handling
Venous blood (10 mL) was collected from each patient between 0900 and 1100 into a plain plastic tube, allowed to clot, and centrifuged at 1000g for 10 min at 4°C. The serum was divided into 0.4-mL aliquots and stored at –20°C for ≤8 weeks until assay. A fresh aliquot was used for each set of measurements and was thawed just before analysis.

Assays
The eight assays used were: BTI Human Osteocalcin RIA kit (Biomedical Technologies, Stoughton, MA); CIS Human Osteocalcin IRMA (ELSA-OSTEO) and RIA (OSTK-PR; both from CIS Biointernational, Gif-sur-Yvette, France); DSL Osteocalcin RIA (Diagnostic Systems Laboratories, Webster, TX); Henning Osteocalcin (Henning, Berlin, Germany); Incastar Osteocalcin RIA (Incastar, Stillwater, MN); Mitsubishi Yuka BGP IRMA kit (IDS, Boldon, Tyne and Wear, UK); and Nichols Institute Human Osteocalcin RIA kit (Nichols Institute Diagnostics). Details of the assay kits are given in Table 1.

Each sample was assayed in duplicate and all kits were used in strict accordance with the manufacturers’ instructions. Samples with OC concentrations greater than the highest standard for any kit were reassayed after appropriate dilution in zero standard solution, as supplied with each assay. There being insufficient tubes in any kit to assay all 48 samples in the same batch, two runs were required for each kit; for each, we used a previously unthawed aliquot of serum.

Statistical Analysis
This was carried out after Delmas et al. (12). The mean OC concentration of the group of nine control subjects was calculated for each assay. Results for each clinical sample with each assay were then divided by the appropriate control mean and expressed as a percentage. An increased OC result was characterized, therefore, by a value of >100%; a value of <100% indicated OC suppression.

Z-scores were calculated for each sample by using the mean and SD for each assay obtained for the control subjects, as follows:

\[
Z\text{-score} = \frac{\text{sample OC} - \text{assay mean OC}}{\text{SD}}
\]

The normal subjects’ OC concentrations were thus normalized to a mean of 0 and an SD of 1 with each assay. Increased and suppressed OC concentrations were therefore characterized by positive and negative Z-scores, respectively. Differences between mean Z-scores for clinical groups and assays were calculated by Student’s t-test.

OC concentrations for each sample were also compared with the reference ranges quoted by the manufacturers, with use of appropriate age- and sex-related intervals where provided.

Recovery Experiment
Synthetic human OC 1-49, provided in 25-ng aliquots by the Nichols Institute, was diluted in the zero calibrator of each assay kit to give final concentrations of 4.31, 2.16, 1.08, 0.54, and 0.27 nmol/L. We assayed these dilutions as described for the clinical samples, using each manufacturer’s standard curve.

Results
Within-batch assay CVs were calculated from patients’ duplicates and are given in Table 1. For between-batch imprecision we used the manufacturers’ figures for the control material provided with each kit, because we made only two runs of each assay. The results for the nine control subjects are shown in Table 2. Different assays gave widely different ranges of values for healthy

### Table 1. Characteristics of the assay kits.

<table>
<thead>
<tr>
<th>Manuf.</th>
<th>Type</th>
<th>Antibody</th>
<th>Label</th>
<th>Std.</th>
<th>Separation</th>
<th>Incubation, h</th>
<th>Sample, μL</th>
<th>Intrassay CV, %</th>
</tr>
</thead>
<tbody>
<tr>
<td>BTI</td>
<td>RIA</td>
<td>Rabbit anti-human OC (C-terminal)</td>
<td>125I-human OC</td>
<td>H</td>
<td>16–20</td>
<td>50</td>
<td>5.4</td>
<td></td>
</tr>
<tr>
<td>CIS</td>
<td>RIA</td>
<td>Anti-human OC, mAb</td>
<td>125I-anti-human OC mAb (N-terminal)</td>
<td>H</td>
<td>2</td>
<td>50</td>
<td>3.0</td>
<td></td>
</tr>
<tr>
<td>CIS</td>
<td>RIA</td>
<td>Rabbit anti-bovine OC</td>
<td>125I-bovine OC</td>
<td>B</td>
<td>20–24</td>
<td>50</td>
<td>5.5</td>
<td></td>
</tr>
<tr>
<td>CIS</td>
<td>RIA</td>
<td>Anti-bovine OC, mAb</td>
<td>125I-bovine OC</td>
<td>B</td>
<td>1</td>
<td>50</td>
<td>4.6</td>
<td></td>
</tr>
<tr>
<td>CIS</td>
<td>RIA</td>
<td>Sheep anti-human OC36–49</td>
<td>125I-OC36–49</td>
<td>H</td>
<td>20–24</td>
<td>50</td>
<td>5.4</td>
<td></td>
</tr>
<tr>
<td>Incastar</td>
<td>RIA</td>
<td>Rabbit anti-bovine OC</td>
<td>125I-bovine OC</td>
<td>B</td>
<td>16–24</td>
<td>50</td>
<td>5.3</td>
<td></td>
</tr>
<tr>
<td>Mitsubishi</td>
<td>IRMA</td>
<td>Anti-human OC34–49, mouse mAb</td>
<td>125I-anti-human OC mAb</td>
<td>S</td>
<td>3</td>
<td>25</td>
<td>5.8</td>
<td></td>
</tr>
<tr>
<td>Nichols</td>
<td>RIA</td>
<td>Rabbit anti-human C-terminal OC</td>
<td>125I-synthetic human OC</td>
<td>S</td>
<td>3</td>
<td>50</td>
<td>5.8</td>
<td></td>
</tr>
</tbody>
</table>

*Human, bovine, or synthetic human. Ab and mAb, antibody and monoclonal antibody, respectively.

CLINICAL CHEMISTRY, Vol. 40, No. 3, 1994 359
subjects. For one individual, measured OC varied from 0.3 to 4.5 nmol/L. However, our results compared well with the manufacturers' quoted reference ranges for their kits. We did observe some consistency in relation to the ranking of individual samples across the assays, but this was not perfect.

Results for the PHPT, Paget disease, steroid-treated, and pregnant groups are summarized in Fig. 1. The mean Z-scores for each assay and clinical group are shown in Table 3. In subjects with PHPT, OC concentrations were above normal with most assays. Only 2 of 48 results were <100% of the control group mean, and these two involved different assays and patients. After excluding these results the overall range was 102–257%. The mean Z-score was significantly greater than zero for all assays except the Incstar kit. The Incstar assay gave results significantly lower than those by the Henning, CIS-IRMA, DSL, BTI, and Nichols kits. Compared with the manufacturers' reference ranges, the BTI and both CIS assays indicated greater OC concentrations for four of five remaining PHPT patients, whereas the Incstar kit classified four OC concentrations as normal and one as low.

In the three patients with Paget disease, serum OC was 149% to 695% of the control group mean. The mean Z-scores were not significantly greater than zero for any assay. When compared with the manufacturers' reference ranges, four kits indicated that all three patients had increased OC; the other four assays indicated that one sample was within their reference limits. Serum OC was <100% by all assays in a fourth patient, who had received a bisphosphonate infusion (not shown in Fig. 1).

Serum OC was <100% of the control group mean in all assays for four of five pregnant subjects; the fifth had values of 110–125% in four of the eight assays. Five of the assays gave a mean Z-score significantly less than zero. The remaining three kits (Henning, Mitsubishi, and CIS-IRMA) all gave significantly higher mean Z-scores than the CIS-RIA and Nichols assays. For four kits, all pregnant subjects had OC within the reference ranges, three kits indicated low OC for four of five patients, and a single assay (Henning) gave a low OC result for one patient.

All five patients on glucocorticoids showed suppression of OC to <100% of the control group mean with all eight assays, although the degree of suppression for any single patient varied among the kits. The most extreme example was one individual whose results ranged from 30% to 96% of the control group mean. All eight kits gave mean Z-scores significantly less than zero. The Incstar assay gave a mean Z-score significantly different from the BTI and CIS-RIA kits. No kit gave results below its stated reference range for all patients, although the Mitsubishi assay did for four of the five. According to the CIS RIA, four of the patients had serum OC within the reference range.

Large between-kit differences were seen in the osteoporosis group (Fig. 2, top). The Mitsubishi kit gave results <100% of the control group mean in 9 of 10 patients, whereas the BTI kit gave results >100% in 8 of the 10. The other assays showed a spread of results around the 100% line; e.g., the Nichols assay gave a range of 34% to 244%. Similarly, 8 of 10 patients had Z-scores that were either above or below zero, depending on the assay used. The rank order of results was consistent for neither patients nor assays. However, all assays did give consistently the lowest value for patient OP8 (Fig. 2). No kit gave a mean Z-score significantly different from zero. However, the mean Z-score for the Mitsubishi assay was significantly less than those for the Incstar, BTI, DSL, and both CIS kits. There was little agreement when the samples were compared with the quoted reference ranges. Three kits indicated normal OC concentrations for all 10 patients. The CIS-RIA and BTI assays indicated raised OC in 6 and 3 of the 10 patients, respectively. The Incstar and Mitsubishi kits both classified one patient (OP8) as having a low serum OC, whereas two patients (OP8, OP9) were so classified by the Henning assay.

The renal failure group showed by far the largest between-kit variation (see Fig. 2, bottom; note logarithmic scale). Except for one patient (R10), whose OC was <100%
of the control group mean by six kits, all the results were increased. Some were extremely high—up to 22-fold the control mean in one case—although for the same patient (R1) another kit gave an increase of only threefold; the corresponding Z-scores were 95 and 5. There was no consistent agreement between apparently similar kit configurations. The CIS two-site IRMA kit tended to give the highest results, but the Mitsubishi kit, also a two-site IRMA, tended to give the lowest. There was no consistency in the rank order of results with different kits. Nor was there good agreement between the assays in terms of their reference ranges. Nine patients had above-normal OC when measured by the BTI, Incstar, and both CIS kits, whereas only 6 of 10 results were above normal by the Henning assay. The Mitsubishi kit was the only one to indicate a low OC in renal failure, albeit in only a single patient (R10). All eight assays gave mean Z-scores for this group that were significantly greater than zero. Again, there were some significant between-kit differences in Z-scores. The mean for the CIS-IRMA was higher than that for the CIS-RIA, Incstar, and Mitsubishi kits. The Mitsubishi kit also gave a mean Z-score significantly lower than that of the DSL, BTI, and Nichols assays.

The results of the recovery experiment (Table 4) indicate major differences in assay standardization, with the Mitsubishi kit alone giving recovery close to that expected. The Incstar and DSL kits gave the lowest recoveries at all dilutions of intact OC: These two assays were very similar in design—using a common antibody source (rabbit anti-bovine OC), label (bovine OC), standard (bovine), and separation (goat anti-rabbit antibody).

Discussion
The clinical states chosen for this study were selected as conditions where serum OC is reported as deviating from normal. PHPT (1), Paget disease (13), renal failure (14), and some cases of osteoporosis (3) are associated with increased OC concentrations, whereas OC is low in pregnancy (15, 16) and glucocorticoid-treated subjects (8-11). The results we obtained were in agreement with previous reports but clearly show that large differences exist between commercial assays.

Although OC is accepted as being a marker of bone formation, its serum concentration is also increased in predominantly bone-resorbing states so long as the two processes are coupled (12). Thus, OC is increased in PHPT (1), as confirmed here, with reasonable agreement among kits. In Paget disease, the essence of the disorder is uncoupling of bone formation and resorption.
which perhaps explains why OC is neither a sensitive test for the presence of Paget disease (13) nor a useful index of its severity or response to treatment (17). We found OC greater than that of control subjects for all three patients studied with all eight kits. Although the results, in terms of percent deviation from controls, were similar for one patient, in the remaining two they were fairly widely dispersed (between 250% and 700%). This may reflect different stages in the natural history of the disease, including the possible release of immunoreactive fragments from resorbed bone.

In pregnancy, OC has been reported as low (15, 18) or even undetectable (16) in serum during the second trimester, and low (16, 18) or normal (15) in the third. We found serum OC to be low in four women in the third trimester but mildly increased in half of the assays in a fifth patient, compared with the control group mean. A suggested mechanism for the decrease in OC during pregnancy is placental-dependent enzymatic degradation (16). If correct, this might result in detectable OC fragments in serum, and may explain both the reported differences between studies and the between-assay differences shown here.

Our finding of suppressed OC in patients on steroid therapy, compared with healthy controls, is in agreement with other reports (8–11, 18, 19), although the degree of suppression seen varied between assays. OC decreases rapidly in response to glucocorticoids, whether they are given orally (8, 19), intravenously (9), or by inhalation (20, 21), and has been proposed as a useful measure of bone suppression (10, 11). Clearly, there will be problems with this diagnostic approach if, as we found, one assay indicates that OC is 96% of normal, while another indicates 30%.

In osteoporosis, Brown et al. (3) found that serum OC correlated well with histological markers of bone formation rate. Serum OC has also been reported as being predictive of the rate of bone loss after menopause (4, 6), as a tool for selecting the appropriate treatment (5, 6), and as a measure of the response to estrogen replacement therapy (4). Instead, we urge caution in attempting to use measurement of OC by uncharacterized methods in the study or management of osteoporosis. A wide range of OC values may be obtained with different kits for the same patient, even when results are expressed as multiples of control means. This will cause greatest diagnostic confusion when results with different kits for an individual subject straddle the 100% line (or the zero line, if Z-scores are used), the situation encountered in 7 of the 10 patients we studied (Fig. 2, top). How would one classify or treat a patient whose OC was 95% of controls with one assay but 250% with another? The situation is not clarified by the use of manufacturers’ reference ranges: One assay indicated that 8 of the 10 patients had above-normal serum OC, while another classified 9 as normal and 1 as low.

OC concentration in plasma increases during renal failure, both directly, due to impaired clearance, and indirectly, due to renal osteodystrophy. OC does not exceed the upper limit of the reference range unless the glomerular filtration rate is <30 mL/min (14); therefore, in mild renal failure OC reflects bone formation alone. In patients undergoing hemodialysis, OC is markedly increased; however, detection of bone disease in the presence of renal failure is still possible. Patients with histologically low-turnover osteodystrophy have mean OC seven to eight times the mean for normal subjects, whereas those with high-turnover osteodystrophy have a 100-fold mean increase in OC (7). Such multiples must be assumed to be specific for the assay used, given the wide variation we have demonstrated in the multiples of the control mean for a given patient. Any study of renal osteodystrophy done with one of the kits evaluated here would, therefore, need to establish new multiples, validated against the same histological markers. Although OC is assumed to be a specific marker of bone formation, in renal failure bone resorption may make a significant contribution to the plasma concentration. Evidence indicates that at least five additional immunoreactive species are present in renal failure patients with high osteoclastic resorption rates, presumably C-terminal fragments (22). The two kits that claim C-terminal specificity (BTI, Henning) are unlikely to have any advantage, therefore, in the presence of renal failure. Indeed, they show no better agreement with each other than with the remaining kits. Nor do the two-site IRMA kits (CIS, Mitsubishi), which might be expected to be less sensitive to fragments, show good agreement. This is reflected in the assays’ variable recoveries of intact human OC. Delmas et al. claimed that the agreement between in-house assays was very good when expressed as percent deviation from normal (12, 23). Their study included only one patient with renal osteodystrophy. However, like us, they presented results on a logarithmic scale; when this is considered, their range of results was similarly wide.

Given the differences in assay type (RIA, IRMA),

---

### Table 4. Analytical recovery of synthetic human osteocalcin (nmol/L) added to zero calibrator of each assay.

<table>
<thead>
<tr>
<th>Added</th>
<th>BTI</th>
<th>CIS-IRMA</th>
<th>CIS-RIA</th>
<th>DSL</th>
<th>Henning</th>
<th>Incatari</th>
<th>Mitsubishi</th>
<th>Nichols</th>
</tr>
</thead>
<tbody>
<tr>
<td>4.31</td>
<td>2.66</td>
<td>3.07</td>
<td>2.22</td>
<td>0.90</td>
<td>4.48</td>
<td>0.79</td>
<td>4.81</td>
<td>1.38</td>
</tr>
<tr>
<td>2.16</td>
<td>0.45</td>
<td>0.62</td>
<td>0.69</td>
<td>0.36</td>
<td>0.67</td>
<td>0.26</td>
<td>1.91</td>
<td>0.47</td>
</tr>
<tr>
<td>1.08</td>
<td>&lt;0.52</td>
<td>0.31</td>
<td>0.33</td>
<td>0.10</td>
<td>0.34</td>
<td>0.17</td>
<td>1.03</td>
<td>0.24</td>
</tr>
<tr>
<td>0.54</td>
<td>&lt;0.52</td>
<td>0.32</td>
<td>0.17</td>
<td>0.03</td>
<td>0.31</td>
<td>0.09</td>
<td>0.59</td>
<td>0.12</td>
</tr>
<tr>
<td>0.27</td>
<td>&lt;0.52</td>
<td>0.10</td>
<td>0.09</td>
<td>0.02</td>
<td>&lt;0.17</td>
<td>0.07</td>
<td>0.31</td>
<td>0.03</td>
</tr>
</tbody>
</table>

n = 1 (mean of duplicates).
standardization (bovine, human), antibody type (monoclonal, polyclonal), and specificity (human, bovine, whole molecule, C-terminal), it is perhaps unsurprising that there is such a wide variation in the absolute OC values determined by different kits. As with in-house assays, results cannot be interchanged between kits. Of greater concern is that the differences were only partially reduced when the data were normalized by being expressed as percentage deviations from normal or as Z-scores. These findings suggest that different assays measure different molecular species, even in the absence of renal failure. This problem is unlikely to be resolved by the use of a common reference material, as has already been demonstrated by Delmas et al. (12).

Recent evidence for the immunological heterogeneity of intact OC comes from experiments with two-site assays with monoclonal antibodies (24). Two different assays gave results that differed by more than twofold, although each measured the intact molecule.

Another possible variable is the calcium-dependence of the assays (25, 26). The conformation of the OC molecule, and hence its immunoreactivity, is critically dependent on the concentration of calcium ions in solution. Some assays are more calcium-dependent than others (26); however, none of the manufacturers indicated the calcium concentrations of their assay buffers. Calcium dependence may, in part at least, account for the differing performances of the kits in the recovery experiment. We agree with Tracy et al. (25) that OC assays cannot be standardized until the differences between antibodies—in terms of epitope specificity and reactivity with multiple forms in serum—are addressed. It also remains to be shown which of these multiple forms are diagnostically altered in which disease states. Nevertheless, new assays continue to proliferate (18, 26) despite the lack of standardization.

Although several studies have shown the potential utility of OC as a marker of bone formation, in some clinical situations this potential will not be fully realized until assays—both commercial and in-house—give concordant results. While the current uncertainty exists we cannot recommend serum OC as a useful measurement in a routine clinical laboratory. Efforts to improve its usefulness would be better directed at issues of standardization than at devising yet more assays.

This work was supported by Nichols Institute Diagnostics, Geneva, Switzerland. We are grateful to Marie Louise Wippermann and Peter Haima for their helpful comments.

References