Journal of Clinical Densitometry: Assessment of Skeletal Health, vol. 12, no. 2, 162–169, 2009
© Copyright 2008 by The International Society for Clinical Densitometry
1094-6950/09/12:162–169/$36.00
DOI: 10.1016/j.jocd.2008.12.005

Original Article

Analysis of ISCD-NIST Survey for Bone Health

Andrew Dienstfrey,[1,]* Tammy Oreskovic,[1] Herbert Bennett,[2] and Lawrence Hudson[2]

[1]Mathematics and Computational Sciences Division of the National Institute of Standards and Technology, Boulder, CO, USA; and [2]Materials Reliability Division, Semiconductor Electronics Division and the Ionizing Radiation Division of the National Institute of Standards and Technology, Gaithersburg, MD, USA

Abstract

In 2007, the National Institute of Standards and Technology and the International Society for Clinical Densitometry designed a survey to prioritize 7 research and standardization action items to improve accuracy and cross-comparability of dual-energy X-ray absorptiometry (DXA) measurements of bone mineral density. In this article, we analyze the 1074 survey responses as one means to define consensus priorities of the community studying bone health and to determine possible correlations between prioritization and demographic data, including geographic location, years of experience practicing DXA, and medical specialty. We find that the distribution of ranks from all respondents is such that we can conclude with statistical confidence that there are perceived distinctions between the relative merits of the 7 action items. Applying a standard vote-counting rule to the data, we determine a complete ranking of the action items. We observe that a consistent ranking of each action item across all demographic subcategories is hard to achieve. When we arrange the 7 action items into 4 groups, however, we can determine a reasonably consistent prioritization. The group containing the development of standard reference databases and phantoms receives the highest priority. In addition, we report consistent themes that emerge from the free-response portion of the survey.

Key Words: Bone mineral density; calibration phantoms; edge detection; reference database; standards.

Introduction

Dual-energy X-ray absorptiometry (DXA) measurement of bone mineral density (BMD) is the primary measurement technology for diagnosing and monitoring bone-related diseases. As with many biomedical imaging technologies, DXA measurement of BMD entails an intricate mixture of physics and image-analysis algorithms. Different DXA scanner manufacturers have arrived at different solutions to these problems. As a consequence, BMD as measured by one DXA scanner does not always agree with BMD measured by another scanner. This is true to some degree within the same manufacturer and scanner model. The problem is significantly larger when trying to compare BMD measurements across manufacturers and clinics *(1)*. This situation not only threatens the credibility of the technique but also stifles the free flow of patients and their medical histories from clinic to clinic.

In light of this situation, in 2006 the National Institute of Standards and Technology (NIST) and the International Society for Clinical Densitometry (ISCD) co-hosted a conference workshop focused on the variability of DXA measurements of BMD. After defining and discussing a list of standards and measurement needs deemed important to increase accuracy and cross-manufacturer consistency of DXA measurements of BMD, the workshop participants ranked the needs in priority order. The results of the workshop activity are reported by Bennett et al *(2)*. Furthermore, roughly 50 persons in attendance at the workshop may not have represented a sufficiently

Received 08/29/08; Revised 12/11/08; Accepted 12/12/08.

*Address correspondence to: Andrew Dienstfrey, PhD, National Institute of Standards and Technology, Math and Computational Sciences Division, 325 Broadway, Mail Stop 891.01, Boulder, CO 80305-3328. E-mail: andrew.dienstfrey@nist.gov

broad cross section of the community interested in questions of bone health. We further refined the list appearing in reference 2 to minimize overlap and redundancies and obtained the following 7 ''action items'' for development in the survey:

- Development of phantoms that validate accuracy at all DXA scan sites and that establish measures of BMD in units of g/cm^2 and g/cm^3 that are traceable to the International System of Units (SI units) [Phan].
- Standardization of edge-finding algorithms and their performance with respect to different soft tissue and density conditions [Alg].
- Standardization of region of interest for all axial DXA scan sites [ROI].
- Development of more complete reference databases for the purpose of consistent evaluations of T- and Z-scores [DB].
- Standardization across manufacturers of quality assurance/quality control protocols for assessing drifts in calibrations [QA/QC].
- Standardization of content and format for DXA reports to enable comparisons among equipment models and manufacturers [Report].
- Standardization, ROIs, reference data, and the like for all peripheral densitometric scan sites [Periph].

The items in square brackets [...] denote the abbreviations that we use for each of the 7 action items in our discussion below. We designed an online survey to determine the priority rankings of the above needs as assessed by a broad cross section of the community studying bone health.

In the following sections we discuss the results of this survey. In the first section, we summarize the survey structure and its online distribution. We next discuss the demographics of survey respondents. After this, we present our analysis of the ranking data obtained from the survey. This analysis is the primary result of the survey. We find that it is hard to assess confidence in determining the precise rank importance of each action item. As an alternative, we propose a coarser analysis in which we group the foregoing 7 action items as follows: Group 1 contains database and phantom development and has the highest priority; Group 2 consists of ROI and QA/QC protocols and has the next level of priority; Group 3 contains report and Algorithm development and has the 3rd level of priority; and Group 4 contains Peripheral standardization with the lowest priority. We find that this 4-group prioritization scheme is reasonably consistent across survey demographic categories. In the last section, we define our coding of the free-responses that we received and present the results of this analysis. Finally, we conclude with a discussion of our major findings and thoughts on future work.

We emphasize that not all sources of inconsistent BMD measurements results derive from issues related to standardization and metrology. It is well documented that inconsistent positioning of a patient being scanned, for example, can produce largely different results. However, because various components of the error analysis for DXA BMD measurements remain to be quantified in a comprehensive and systematic manner, now is the time to set priorities with the goal to minimize sources of error. The survey described in the following section is one means of approaching this task.

Survey Structure

The survey opened with a series of demographic questions:

- In which country do you primarily work?
- Check all specialty societies of which you are currently a member—American Association of Clinical Endocrinologists (AACE), American College of Obstetricians and Gynecologists (ACOG), American College of Radiology (ACRad), American College of Rheumatology (ACRh), American Society for Bone and Mineral Research (ASBMR), International Society for Clinical Densitometry (ISCD), North American Menopause Society (NAMS), and The Endocrine Society (TES).
- What is your specialty?
- How many years have you practiced your specialty?
- How many years of experience do you have reading DXA?

These demographic questions were followed by a brief discussion of the motivation and goals. Three guiding assumptions were stated as follows:

1. DXA is the primary measurement technology for diagnosing, monitoring, and ultimately contributing to bone health.
2. The accuracy of BMD as measured by DXA scans is not optimal for all of its intended purposes.
3. Accuracy of DXA could be significantly enhanced through standards, measurements, and compliance efforts.

Survey respondents were asked whether they believed that these 3 assumptions were appropriate and complete. A negative response was followed by an opportunity to indicate why. After this, the list of 7 action items indicated earlier was presented. Respondents were asked to prioritize the action items in rank order from 1 to 7 with no ties permitted. In this ranking, 1 indicates the item that is most significant for improving accuracy and cross-comparability of DXA—measured BMD, and 7 is the least significant item. The final survey question allowed for comments and free responses.

The survey was online for 6 wk (2007-10-23 through 2007-12-04) at the ISCD's Web site. Invitations to complete the survey were distributed by e-mail to the combined membership lists of the 8 stakeholder societies concerned with bone health listed earlier. In total, 1074 respondents who are members of one or more of these societies volunteered to complete the survey. Here a *complete* response is defined as a response for which the ranking question was completed, that is, if the ranking question was completed the ranks were counted even if demographic questions for the same respondent were unanswered or invalid. We restrict analysis to these complete responses. Most of the responses, about 90%, came from the United States. However, there was international participation including respondents self-reporting as practicing in most regions of the world, for example, Canada, Europe, and Latin America. The respondents included people from several medical disciplines, for example, radiologists, endocrinologists, rheumatologists, and so forth, and they self-reported a wide range in number of years of experience in working with this technology from 0 to 20 yr.

Survey Demographics

Survey respondents represented a broad cross section of the community studying bone health. This is summarized in Table 1. The respondents self-reported as practicing in countries representing most large geographic areas, including the Arabian Peninsula. Even though most respondents report more than 10 yr experience reading DXA, we find that years of experience had little effect on the ranking of priorities. The ISCD is the largest single society represented. However, memberships of other societies were represented as well. We note that respondents were allowed to check as many societal memberships as needed; hence this column sum exceeds the number of overall survey responses. Finally, a broad cross section of medical specialties participated in the survey. We do not attempt to draw inferences about any of the demographic subcategories as such. For example, we do not attempt to weight subcategories by response rate to achieve a consistent weighting in the consensus average. Rather, survey respondents are a self-selected group with interests and opinions for improving DXA standards and measurements; their demographic data are used only for categorical purposes.

Rank Analysis

We are interested in the analysis of the action item rankings, as one of the primary goals of the survey was to determine a consensus prioritization among these items. Survey respondents were required to rank prioritize all 7 action items with no ties allowed. Tallying the ranks result in each action item receiving a distribution of ranks from 1 to 7. These distributions can be considered for all respondents or various demographic subcategories, for example, Canadian respondents, endocrinologists, and others, to look for consistency of priorities across cohorts.

Ordinal Statistics and Concordance

The distribution of votes from all survey respondents is presented in Table 2. The table shows the number of votes each action item received by rank and the median rank and its uncertainty. The final column shows the consensus priority as determined by Borda count (see Arrow *(3)* and later discussion). This ordering is referred to as the "global consensus." in this article. The order of the action items in the table is the order in the list given earlier, which was also the order in which they appear in the online survey. We note that the global consensus order is not the same. For example, Phantom development received 304 rank, 1 vote; 149 rank, 2 votes; and 146 rank, 7 votes. The median rank of the underlying random variable is estimated to be 3 ± 0.19. The consensus is that development of phantoms is 2nd to development of standardized databases as a priority activity for increasing accuracy and cross-comparability of DXA measurement of BMD. The consensus priority order is determined by a traditional weighted scoring technique (Borda count). We defer discussion of this procedure for the following section. Here, we present preliminary statistical analyses. For clarity, we restrict the discussion to results that treat all respondents as a single category. The analyses reported here and in the following were performed for all demographic subcategories also.

In Fig. 1 a graphical representation of Table 2 as a series of histograms is shown. Each of the 7 action items in Fig. 1 has 7 bars associated with it. The 1st bar on the left represents the

Table 1
Demographic Data of Survey Respondents—Respondents Could Type in Any Number as Answer to the "Years Experience" Question

Geographic region	N	Yr	N	Society	N	Specialty	N
United States	936	A: $1 \le yr < 3$	100	AACE	50	Endocrinology	119
Canada	45	B: $3 \le yr < 5$	97	ACOG	64	Family Medicine	54
Europe	34	C: $5 \le yr < 10$	298	ACRad	361	Internal Medicine	59
Asia	14	D: $10 \le yr$	372	ACRh	84	Nuclear Medicine	35
South America	13			ASBMR	151	OBGYN	124
Arab Peninsula	6			ISCD	627	Radiology	406
Other	26			NAMS	99	Rheumatology	119
				TES	88	Women Health	35
Total N	1074		887		1524		951

Answers are grouped into the brackets (A, B, C, and D) as shown. The column sums can be less than 1074 as some respondents chose not to answer all of the demographic questions. The "Society" column sum exceeds 1074 as respondents often belonged to more than 1 society.

Abbr: AACE, American Association of Clinical Endocrinologists; ACOG, American College of Obstetricians and Gynecologists; ACRad, American College of Radiology; ACRh, American College of Rheumatology; ASBMR, American Society for Bone and Mineral Research; ISCD, International Society for Clinical Densitometry; NAMS, North American Menopause Society; TES, The Endocrine Society.

Table 2
Distribution of the Number of Votes and Summary Statistics for Each Action Item

	1	2	3	4	5	6	7	Median	Consensus
Phantoms	304	149	103	135	115	122	146	3 ± 0.19	2
Algorithm	57	172	174	152	185	200	134	4 ± 0.14	6
ROI	127	185	227	189	146	128	72	3 ± 0.14	3
Database	230	179	170	190	127	97	81	3 ± 0.14	1
QA/QC	83	172	199	201	240	126	53	4 ± 0.10	4
Report	132	146	131	111	155	264	135	5 ± 0.19	5
Peripheral	141	71	70	96	106	137	453	6 ± 0.19	7

Survey respondents considered as a single category.
Abbr: ROI, region of interest; QA/QC, quality assurance/quality control.

number of respondents who gave that action item a rank of 1. The next bar corresponds to the number of respondents who gave that action item a rank of 2, and so forth. A rank of 1 indicates the highest priority and a rank of 7 is the lowest.

Median rank and its associated uncertainty is reported in Table 2. Our formula for computing the uncertainty in the median estimate is given in the Appendix. The use of median as a measure of central tendency, as opposed to mean, is more appropriate for ordinal rank data *(4)*.

The box plot (Fig. 2) is a visual representation of Table 2 statistics. Fig. 2 presents the action items in sorted order with the top item, databases considered to be most important. The black vertical lines indicate the median scores. The lateral left and right extents of the boxes correspond to the 1st and 3rd quartiles, respectively. The notches in the box indicate 95% confidence intervals (CIs) for the estimate of the median as computed by Eq. (A1) in the Appendix.

We computed Friedman's statistic to assess the degree of distinction between action items. Our analysis follows Lehmann *(5)* and details are provided in the Appendix. Friedman's statistic is designed to test the null hypothesis,

H_0 = "Voters randomly assigned ranks to the items with equal probability."

In other words, when H_0 is true, then the distribution of votes reflects no discernible preference among action items. To test H_0, we compute Friedman's statistic Q [see Eq. (A2)] and compare the value against the null distribution by way of the confidence p value. One interpretation of the p value in relation to an observed value, Q_{obs}, is that assuming H_0 is true, one would expect a value of Q greater than or equal

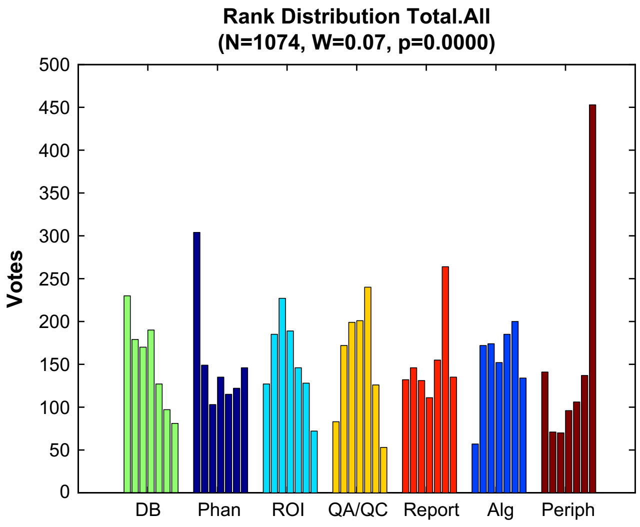

Fig. 1. Distribution of all votes (N = 1074) for the 7 action items in Table 2. For each item the histogram shows the 7 bars. The 1st bar indicates the number of rank 1 votes that the item received; the 2nd bar the number of rank 2 votes, and so forth. The action items are presented in the order determined by the scoring system discussed in the article. In all plots colors are used consistently to identify action items.

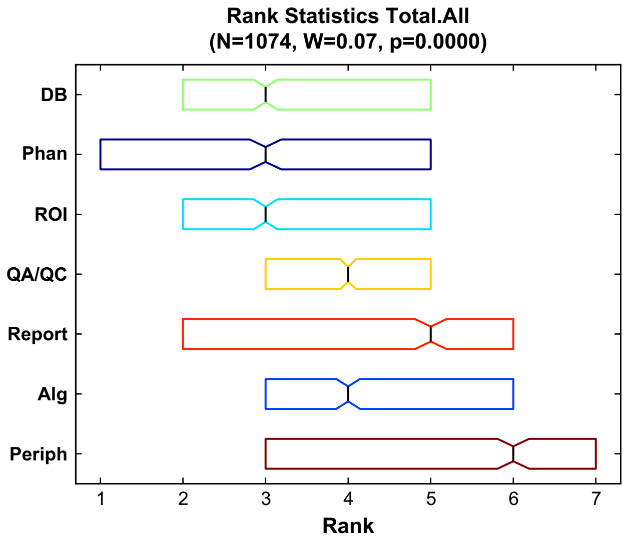

Fig. 2. Box plots showing the summary statistics for Table 2. The short-black vertical lines show the sample median ranks that serve as estimates of the true median. The notches indicate the 95% confidence interval for this estimate. The extent of the boxes indicates the 25% and 75% quartile ranges.

to Q_{obs} with probability p. The p value is computed by use of Eq. (A4). We find that for almost every demographic subcategory—that is, partitioning the respondents by geography, medical specialty, years of experience, and societal affiliation—we can reject H_0 with more than 99% confidence ($p < 0.01$). Fig. 2 shows this conclusion is supported by the observation that the estimates of the median ranks of the action items are such that the 95% CIs [see Eq. (A1)] for all 7 items do not overlap. This lack of overlap is consistent across most all-demographic subcategories, and provides evidence that there are perceived differences among the 7 action items. The exceptions to this were the categories of respondents reporting as practicing in the geographic regions: Arabian Peninsula (N = 6, $p = 0.29$), Asia (N = 14, $p = 0.02$), and South America (N = 13, $p = 0.27$). In these cases the combination of number of respondents, N, and the diversity of responses were such that the rankings could be viewed as the result of random orderings with no preference. As a consequence it is not meaningful to report ranking results from these 3 subcategories.

In summary, although the histogram plots such as that shown in Fig. 1 do not reveal *obvious* structure, with the 3 exceptions noted above, the distributions of ranks suggest that it is unlikely that they were assigned randomly with equal preference to all items. We discuss our technique for determining consensus ranks in the following section.

Rank Prioritization

Aggregating a collection of rankings to determine a consensus rank is a well-known problem in voting and social choice theory. For an early and famous example of the difficulties in the field see Arrow *(3)* and for more modern treatments, see Young and Levenglick *(7)* and Saari *(6)*. At present there are several competing algorithms for the job with no clear "optimal strategy" among them. We selected a traditional positional weighting scheme referred to as a Borda method *(6)*. Applying this procedure to the present survey we assign the following scores: the first place action item on every ballot received a score of 6, the second place item a score of 5, and so forth, until the lowest ranked item on a ballot received a score of 0. The scores are assigned to each ballot individually, and then summed over all ballots within the demographic category of interest. The items are then ranked in descending order by the Borda score, that is, the highest score is the "winner." In short, the Borda score may be understood as a weighted mean with a particular assignment of weights to ballot positions. Given the vote distributions in Table 2, we compute the following Borda scores: database development, 3876; Phantom development, 3738; standardization of ROI, 3582; standardization of quality assurance and quality control 3363; standardization of report format, 2953; standardization of Algorithms, 2924; and standardization of Peripheral DXA technologies, 2118.

We stress that both the choice of a positional scoring method, and subsequently the selection of weights to be applied, can affect the results. For example returning to Table 2, whereas the development of Phantoms clearly receives the most rank 1 votes, the Borda scoring scheme values the relatively large number of 2nd and 3rd place votes received by database development to the extent that the latter edges out the former by a narrow margin. One could construct an alternative weighting scheme that allocates higher value to 1st place ranks relative to the middle than does the arithmetic sequence 6, 5,…, 1, 0. In such cases the consensus prioritization between database and phantom development would reverse in priority.

Fig. 3 presents the consensus rankings for most demographic subcategories. The leftmost column shows the action items in descending priority order from top to bottom. Database development, the highest priority item, is at the top. This is the priority order of the "global consensus" and is the same as the last column of Table 2. Each of the remaining columns in the plot refers to a demographic subcategory. As noted, the leftmost column consists of all respondents. The next 3 columns correspond to partitioning respondents by geography. The next 4 columns refer to years of experience. The next 8 columns show break-outs by societal affiliation. The final 8 columns arrange respondents by medical specialty. These columns may be compared with entries in Table 1; for example, the definitions of A, B, C, and D are contained therein. For each subcategory the corresponding survey responses were scored by the Borda method, and a prioritization was determined. The symbols in the column consistently refer to the same action item, for example, a green square is database development, a dark blue "X" is phantom development, and so on. An empty space within a column indicates that the priority ascribed to the item is consistent with the global consensus. A symbol in a column indicates a point of disagreement

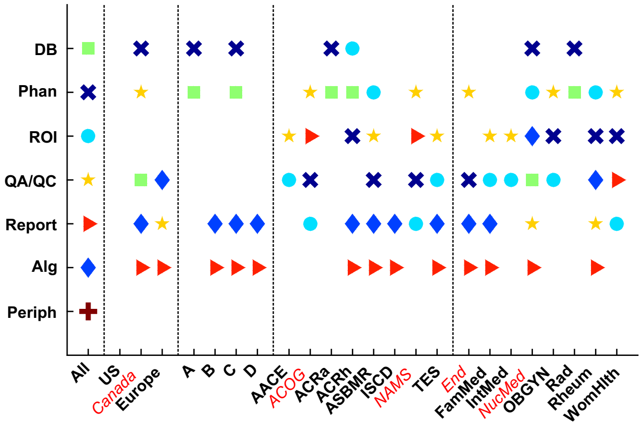

Fig. 3. Consensus rankings by various demographic groups. The leftmost column shows the action items in the priority order as determined by all survey responses, defining a "global consensus." The other columns show the priorities of subgroups. A blank space in the column indicates agreement with the global consensus. A symbol indicates a difference. The column labels typeset in red italics, for example, Canada, indicate a subgroup with a substantially different prioritization.

with the global consensus, in which case the symbol type refers to the action item that was put in that rank instead. For example, the Canadian respondents to the survey agreed with the global prioritization in placing ROI and peripheral standardization activities as 3rd and 7th, respectively. However, phantom development was ranked as the most important followed by QA/QC protocols. The top importance ascribed to these activities breaks with the popular opinion, which has them ranked 2nd and 4th, respectively.

We draw the following conclusions from Fig. 3. With the exception of the US column, all other columns contain reordering of action item ranks. As the US category corresponds to more than 90% of all survey respondents, this perfect agreement with the ‘‘global consensus’’ is likely. The United States aside, the figure shows that the prioritization determined by global consensus is not entirely consistent with priorities defined by subcategories. However, looking more carefully, we find that most of the reorderings are adjacent transpositions. For example, it is common for the orders of report and algorithmic standardization to oscillate between priorities 5 and 6. To a lesser extent, the same could be said for database and phantom development receiving top priorities. This last point is to be anticipated as we noted previously that it would be relatively easy to construct a different weighting scheme that would toggle the order of these 2 items within the global consensus. Allowing for these 2 ambiguities (database and phantoms, and ROI and reports) the consensus appears to be more widespread. For example, outside of these 2 transpositions there are no other differences in which respondents are categorized by years of experience.

In summary, we observe that the final ranking in its every detail is not a very precise affair. However, a slightly coarser grouping suggests itself as being possible and agreeable to all parties. In this re-factoring of action items, database and phantom development are given a combined highest priority, followed by a 2nd group consisting of ROI and QA/QC standardization, report and algorithm development are tied for the next level, and peripheral standardization is the lowest priority.

Free-Response Analysis

We conclude our analysis of the survey with a brief discussion of the comments and free responses. To reiterate, the 3 guiding assumptions underlying the survey were stated and are as follows:

1. DXA is the primary measurement technology for diagnosing, monitoring, and ultimately contributing to bone health.
2. The accuracy of BMD as measured by DXA scans is not optimal for all of its intended purposes.
3. Accuracy of DXA could be significantly enhanced through standards, measurements, and compliance efforts.

The first survey question then asked whether respondents agreed with these assumptions. A large majority of all survey respondents (N = 978 or about 91%) agreed. Those who did not agree were provided the opportunity to comment. Additionally, the final question of the survey, asked for open-ended comments.

We do not distinguish between text responses received from either opportunity to comment, that is, comments on the guiding assumptions or the open-ended comments. Out of the 1074 survey respondents we received 269 text responses representing 244 distinct individuals. Some respondents who chose to enter comments did so at both opportunities. All comments were read first by members of the NIST Bone Health team, and then discussed in a group setting. We discarded some comments that were not relevant to the goals of the survey or inappropriate in some other way. We categorized the remaining comments into 7 major themes. Each member of the NIST team then worked alone to match comments to themes. Longer comments could crosscut these themes and we counted them as many times as they were relevant. We compared the resulting lists and adjusted discrepancies. Table 3 contains the 7 themes and numbers of comments matching them.

From Table 3 we see that the most common theme indicated that human-based errors—for example, patient positioning, technologist training, and certification—were one of the primary causes for the lack of cross-comparability. The ISCD identified this earlier and has ongoing programs to ameliorate this problem. The ‘‘non-technical’’ measurement needs referred to in theme 5 include refinement of databases to improve consistency of reports of T- and Z-scores, the need to develop standards and guidelines for use of DXA in pediatric and male patients, and the need for standardized report formats to aid cross-comparability.

Some 11 people responded with concerns over the recent DXA reimbursement rate cuts. In some cases these responses indicated that the pressure to decrease time reading DXA scans would lead to a concomitant reduction in accuracy. This is

Table 3
Survey Analysis of Free-Response Data Aligned by Theme

Theme	Number of responses
1. DXA suffers from inherent limitations.	27
2. DXA is sufficiently accurate for its intended use.	24
3. Alternative modalities to DXA should be used to measure bone health.	32
4. Measurement inaccuracy is due to human error.	41
5. Measurement inaccuracies are because of nontechnical measurement needs.	35
6. Primary concern is the recent DXA reimbursement rate cut.	11
7. Concerns over the role and value of standards and/or a central body.	27

Abbr: DXA, dual-energy X-ray absorptiometry.

notable in that the survey specifically did not address reimbursement rates as other surveys have been circulated for that task. We think that if the issue of reimbursement were brought explicitly to the attention of respondents, then the response rate indicating such concerns would be significantly higher.

The sizable number of respondents indicating concerns over the role and value of either a standards process or a central standards process, for example, one certified by NIST, is notable. These comments were often combined with concerns about the cost-effectiveness of such approaches. The idea is that increased standardization would take the form of more time spent in clinical settings to calibrate DXA scanners. We share the concern that this is potentially very relevant to DXA practitioners. Our perspective of the problem at this time is that the quantitative aspects of the problem need to be carefully defined before such concerns should be addressed. The scope and shapes of solutions should incorporate cost-benefit analyses. However, such cost-benefit analyses can be performed only after careful error analyses have been constructed and verified. Finally, 2 survey respondents mentioned that the current state of standardization and accuracy of DXA measurement of BMD is similar in many ways to the state of mammography in the early 1990s, before the 1992 *Mammography Quality Standards Act* (MQSA).

Conclusion

In conclusion, the goals of this survey were to determine the extent of consensus from the bone-health community around the following propositions: (1) DXA remains the primary technique for determining bone health, (2) BMD, determined by DXA, is not sufficiently accurate and consistent across machines and clinics, and (3) that standardization activities, including reference data, could make a significant contribution toward optimizing DXA for its intended uses thereby improve patient care. Over 90% of respondents agreed with these assumptions concerning the present state of DXA and its central role in clinical practice. This is not inconsistent with recent developments of fracture-risk models for bone health assessment for which the DXA score may or may not appear as one of several input factors *(8)*.

The final proposition gauges the value of standardization efforts and the related priority ranking of ways to improve the application of DXA. Our conclusion from the analysis presented here is that solutions can be viewed in 4 groups. In order of importance these groups are as follows: (1) database and phantom development, (2) standard definition of regions of interest and improvement of QA/QC protocols, (3) standardization of report format and image-analysis software algorithms, and, (4) development and standardization of peripheral DXA technology. Although the ordering of individual solutions can change by choice of analysis procedure, we find that this grouping of solutions and their ordering largely reflect the consensus of the multifaceted community of stakeholders in bone health sampled by this survey.

This broad-based survey was intended to elicit the views of the bone-health community as to the sources of variability in reported DXA scores and the need to reduce measurement uncertainties. Given the public-health ramifications and widespread concerns for the bone health of an aging society, it is important to identify and quantify the various components of the error analysis as DXA remains a primary measurement technology. In light of the survey results, we conclude that added effort in the area of standard test objects, analytical methods, and reference data would significantly increase the accuracy and reproducibility of results, and for the first time, allow quantification of the errors attributable to other sources, such as positioning of the patient. Such investment would improve patient care, reduce wasted effort and cost, and enhance confidence in the DXA technique for all stakeholders.

Acknowledgments

The authors thank Stephen Dienstfrey, former statistician with the Veterans' Administration, for extensive consultation on analysis of survey data. Additionally, the authors thank Jack Wang, Hung-Kung Liu, and Jim Filliben, all of NIST's Statistical Engineering Division, and Dr. John Shepherd of University of California at San Francisco, for their comments and suggestions on previous versions of this document. We also thank Martin Rotblatt, Jennifer Gentry, and Drs. Neil Binkley, Stephen Petak, Michael Lewiecki, and Lisa Karam for their continued support, and all the officers of the ASMBR, AACE, ACR, NAMS, NOF, and TES, who encouraged their respective members to participate in the DXA Survey.

References

1. U.S. Department of Health and Human Services. 2004 Assessing the risk of bone disease and fracture. Bone Health and Osteoporosis: A Report of the Surgeon General. U.S. Department of Health and Human Services, Rockville, MD, 208. Office of the Surgeon General.
2. Bennett HS, Dienstfrey A, Hudson LT, et al. 2006 Standards and measurements for assessing bone health-workshop report co-sponsored by the International Society for Clinical Densitometry (ISCD) and the National Institute of Standards and Technology (NIST). J Clin Densitom 9:399–405.
3. Arrow KJ. 1963 Social choice and individual values. 2nd ed. Yale University Press. New Haven, CT.
4. Kitchenham B, Pfleeger S. 2003 Principles of survey research part 6: data analysis. ACM SIGSOFT Software Eng Notes 28:24–27.
5. Lehmann EL. 1975 Nonparametrics: statistical methods based on ranks. Holden Day series in probability and statistics. Holden Day, San Francisco, CA.
6. Saari DG. 2001 Decisions and elections: explaining the unexpected. Cambridge University Press. New York, NY.
7. Young HP, Levenglick A. 1978 Consistent extension of condorcets election principle. SIAM J Appl Math 35:285–300.
8. Kanis JA, on behalf of the World Health Organization Scientific Group. 2008 Assessment of osteoporosis at the primary health care level [Technical Report]. World Health Organization, World Health Organization Collaborating Centre for Metabolic Bone Diseases, University of Sheffield, UK.
9. Friedman M. 1940 A comparison of alternative tests of significance for the problem of m rankings. Ann Math Stat 11:86–92.
10. Kendall MG, Smith BB. 1939 The problem of m rankings. Ann Math Stat 10:275–287.

Appendix.

Statistical Formulas

We treat the ranks as an ordinal variable and therefore use the median as an estimate of the central tendency *(4)*. The uncertainty on this estimate is computed as follows: Let r_m be the median rank. The 95% confidence interval (CI) for r_m is $[r_{\text{lower}}, r_{\text{upper}}]$ defined as

$$\begin{aligned} \Delta m &= 1.57(r_3 - r_1)/\sqrt{\mathrm{N}} \\ r_{\text{upper}} &= \min\{r_m + \Delta m, r_3\} \\ r_{\text{lower}} &= \max\{r_m - \Delta m, r_1\} \end{aligned} \tag{A1}$$

where r_3 and r_1 are the 3rd and 1st quartile ranks, and N is the number of respondents being considered. In other words, the CI is symmetric about the median unless the interval extends beyond the quartiles, in which case the quartile rank is used.

Our presentation of Friedman's statistic follows Lehmann *(5)*. As the survey has 7 action items (ie, "treatments") and repeat rankings are not allowed, if one assumes H_0 is true, then the mean action item rank is $(7+1)/2 = 4$. Friedman's statistic is the scaled sum of squared differences,

$$Q = \frac{12N}{7(7+1)} \sum_{s=1}^{7} (\overline{R}_s - [(7+1)/2])^2 \tag{A2}$$

Here N is the number of respondents and $\overline{R}_s$ is the mean of the sth action item. We reject H_0 for large values of Q. Under the normalization (A2), the large N asymptotic distribution for Q is a chi-square variate with 6 degrees of freedom, χ^2_6. For all subcategories we assume that N is sufficiently large that this asymptotic distribution is valid *(9)*. Confidence p values are computed accordingly. In place of Q, for consistency across different size groups, we report Kendall's W

$$W = Q/N(7-1) \tag{A3}$$

This rescaling of Q is such that $0 \leq W \leq 1$. Kendall and Smith *(10)* provide other interpretations of W.

As an example, using the data of Table 2, we compute $Q_{\text{all}} = 443.6$ and the associated $W_{\text{all}} = 0.069$ (N = 1074 for all survey respondents). Using the complementary cumulative distribution function of a χ^2_6 random variable, the probability of observing $Q \geq Q_{\text{all}}$ when H_0 is true is computed by,

$$p_{\text{all}} = 1 - F_{\chi^2_6}(Q_{\text{all}}) = 0 \tag{A4}$$

In this example $Q_{\text{all}} = 443.6$ is sufficiently large that p_{all} is effectively 0. In conclusion, because the probability of observing Q_{all} (or higher) when H_0 is true is extremely small, we may then assert that H_0 is false.