Int J Legal Med (2007) 121:477–482
DOI 10.1007/s00414-007-0199-x

ORIGINAL ARTICLE

Accuracy of metric sex analysis of skeletal remains using Fordisc® based on a recent skull collection

F. Ramsthaler · K. Kreutz · M. A. Verhoff

Received: 10 October 2006 / Accepted: 20 August 2007 / Published online: 25 September 2007

Abstract It has been generally accepted in skeletal sex determination that the use of metric methods is limited due to the population dependence of the multivariate algorithms. The aim of the study was to verify the applicability of software-based sex estimations outside the reference population group for which discriminant equations have been developed. We examined 98 skulls from recent forensic cases of known age, sex, and Caucasian ancestry from cranium collections in Frankfurt and Mainz (Germany) to determine the accuracy of sex determination using the statistical software solution Fordisc® which derives its database and functions from the US American Forensic Database. In a comparison between metric analysis using Fordisc® and morphological determination of sex, average accuracy for both sexes was 86 vs 94%, respectively, and males were identified more accurately than females. The ratio of the true test result rate to the false test result rate was not statistically different for the two methodological approaches at a significance level of 0.05 but was statistically different at a level of 0.10 (p=0.06). Possible explanations for this difference comprise different ancestry, age distribution, and socio-economic status compared to the Fordisc® reference sample. It is likely that a discriminant function analysis on the basis of more similar European reference samples will lead to more valid and reliable sexing results. The use of Fordisc® as a single method for the estimation of sex of recent skeletal remains in Europe cannot be recommended without additional morphological assessment and without a built-in software update based on modern European reference samples.

Keywords Metric analysis · Forensic database · Sex determination · Fordisc® · Forensic anthropology · Population dependence · Skeletal remains

F. Ramsthaler (✉)
Department of Legal Medicine, University of Frankfurt/Main,
Kennedyallee 104,
60596 Frankfurt/Main, Germany
e-mail: ramsthaler@em.uni-frankfurt.de

K. Kreutz
Institute of Forensic Anthropology,
Kronzenborner Weg 14,
35435 Wettenberg, Germany

K. Kreutz
Institute of Biology and Chemistry, University of Hildesheim,
Marienburger Platz 22,
31141 Hildesheim, Germany

M. A. Verhoff
Department of Legal Medicine, University of Gießen,
Frankfurter Straße 58,
35392 Gießen, Germany

Introduction

Morphological trait analyses for sexing skeletons are well established, but they may be difficult due to breakage or postmortem damage of the skeletal remains. Since there is often a size difference between the two sexes, different studies have attempted to differentiate males and females on the basis of bone measurements.

In discriminant function analysis, reliable equations have been developed for different bones or bone fragments which are widely used today in anthropological research [4, 8, 9, 15, 16]. Most authors repeatedly emphasize the fact that the use of any sexing method should be limited to identical or very similar populations [18, 29, 42]. Secular changes and population-specific differences among modern, mobile societies complicate the implementation of used

methods which have, however, only rarely been developed from recent material [5, 20, 21, 23, 25, 26, 31, 35, 39, 46].

Nonautosomal nuclear DNA analysis based on amelogenin gene identification is established for sexing bones, but DNA is not recoverable from remains in all circumstances [17, 44, 45]. Chang et al. [6] reported about a high failure rate in some populations; in previous studies it was found that approximately two thirds of all samples tested gave positive results in DNA-based sex determination [7, 11, 12, 13, 27, 28]. Currently, because history, post mortem interval, and ancestry is often unknown at the beginning of the anthropological investigation and because of the risk of contamination, DNA extraction and amplification from skeletal material do not—contrary to expectations—make traditional anthropological sexing methods dispensable [24, 38, 44].

In 1993, Jantz and Ousley [22] first publicized the osteological software Fordisc®. The software package enables a computerized analyses of sex, ancestry, and body height through flexible biometric approaches based on the availability of different measurements from skeletons [37]. At present, the reference database is based on more than 1,200 cases from the American Forensic Data Bank (FDB) and from the Terry and Hamann Todd Collection which contains more than 1,700 reference cases from the early nineteenth century [10, 19, 30, 31, 37].

Fordisc®, which currently exists in the version 3.0, enables sex determination by computing variable biostatistical functions based on a maximum of 36 different cranial measurements. The fast, efficient, and straightforward usage of the software for different discovery conditions has contributed to its increasing success throughout the USA and worldwide [41, 43]. It is not only of importance that the FDB represents recent forensic cases but also that the data pool is continually being expanded by verified, new, well-documented forensic cases that are being passed on to the developers [22, 34, 36, 37]. This renewed data basis leads to a continued recalculation of the basis statistics.

On the basis of studies on a Spanish bone collection from the sixteenth and seventeenth centuries, Ubelaker [43] reported strongly differing results between the morphological method and the use of the Fordisc® software in regard to sex and population affiliation. One of the goals of this study was to test the effect of the transfer of discriminant function formulae developed on forensic collections in North America to osteometric circumstances in present-day Europe.

The dispute about which methodical approach is better has recently flagged since the majority of scientists share the conviction that the combined use of metrical and morphological methods is the best way to determine osteological profiles [14, 38]. Both the high accuracy and verifiability of linear measurements and the advantages of a detailed and diverse morphognosis based on the examiner's high professional qualification and experience can be drawn on by integrating both approaches.

To compare the accuracy of Fordisc® results with morphological standards of sex determination was the second aim of this work.

Materials and methods

The Centers of Forensic Medicine at Frankfurt and Mainz University are curators of a recent skull collection of more than 100 forensic cases from the past 40 years. The collections represent individuals who lost their lives through unnatural causes of death.

Preparation and maceration was necessary to perform further examinations of gunshot wounds and blunt or sharp force injuries that were ordered by the attorney's office or judge. Scientific analysis of this anonymized material has

Table 1 Example of sex estimation by using a scoring system of nine nonmetric sexual traits

Nonmetric sexual traits	Weight	Distinctness	Product	Typical female	Typical male
	w	*x*	*xw*	(−) minus	(+) plus
Skull size and shape	2	+1	+2	Small, round −	Large, square +
Supraorbital ridge	2	−1	−2	Flat −	Prominent +
Upper orbital rim	2	+1	+2	Sharp −	Blunt +
Chin shape	1	+2	+2	Flat, rounded −	Square +
Tubera frontalia	1	0	0	Distinctive −	None +
Mastoid process	2	+2	+4	Gracile −	Large, robust +
Occipital muscle ridges	1	+2	+2	Slight −	Strong +
Zygomatic arch extension	1	+1	+1	None −	Extends to the auditory meatus +
Mandible ramus	1	−1	−1	Narrow, less angled −	Wide, sharply angled +
		Sum	+10		

Any single score is calculated as a product of weight (*w*) and distinctness (*x*), with *x* values ranging from −2 to +2. If the observed trait is indifferent, *x*=0. Sex determination can be estimated by calculating a sum of all nine single scores. A sum >1=male, a sum <1=female.
x (*−2* hyperfeminine; *−1* feminine; *0* indifferent; *+1* masculine; *+2* hypermasculine)

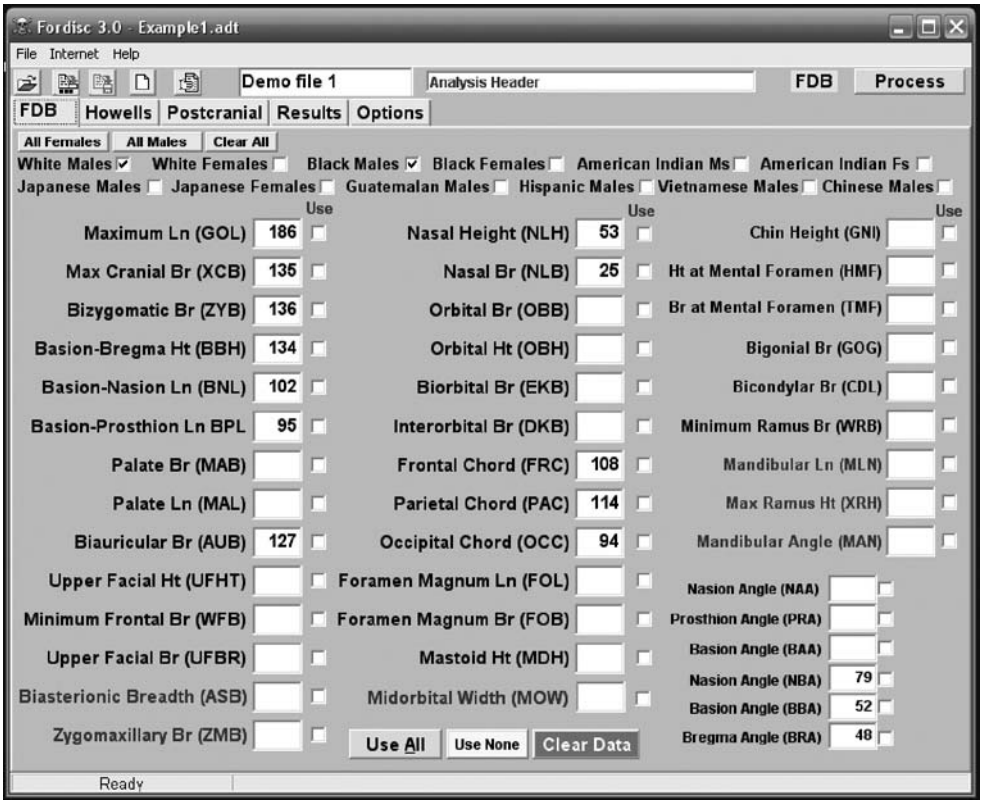

Fig. 1 Software Fordisc® (vers. 3.0) input mask

been approved by the ethical board of the University of Frankfurt (Project number FOR 1A).

Concerning these two collections, which represent a Middle European population from 1968 to 2005, we have knowledge both of the sex and age. All cases included in this study have been clearly identified; there are no skulls where sex has been determined by postmortem analyses.

From the collections, 98 skeletonized, well preserved crania (65 male, 33 female) were initially examined morphologically. For each case, a score system was used to calculate a degree of sexualization (Table 1). Table 1 lists the morphological characteristics and shows the calculation method introduced by Acsádi and Nemeskéri [1] by which the sex determination was accomplished.

The metric measurements for all 98 crania were collected in a standardized form and analyzed by Fordisc® vers.3.0 discriminant functions (Fig. 1). The cranial and mandibular landmarks were defined and recorded very strictly to ensure the accuracy of the measurement procedures. This standardization of data collection, as it is used by Fordisc®, was very important in order to guarantee comparability.

The skulls to be measured in the first and second run of the examination were chosen in a masked, random sequence in order to avoid an impact on examiner bias. Statistical analysis was done with a cross table test to compare true classification rates for both methods (Table 2). In addition, the true and false classification rates for males and females for each method are shown separately in the 2×2 Tables (Tables 3 and 4).

Receiver operating characteristic curve analysis (ROC) was used to compare the diagnostic performance of the two

Table 2 Cross table showing true and false classification rates for both methods

	True	False	Total
Fordisc®	84	14	98
Morphological	92	6	98
Total	176	20	196

Statistics: n=98. Degrees of freedom, 1. χ^2=3.56 (no Yates' correction for continuity). P value is two-sided (or two-tailed). For significance at the 0.05 level, χ^2 should be greater than or equal to 3.84. The difference between the two test methods is significant at level 0.10 (p=0.06).

Table 3 2×2 Table showing true and false classification results for males and females using Fordisc®

	Real male	Real female	Sum
Fordisc® male	58	7	65
Fordisc® female	7	26	33
	65	33	98

Both sexes can be estimated with about 86% accuracy. Sensitivity= 0.89; specificity=0.79; n=98.

Table 4 2×2 Table showing true and false classification results for males and females using morphognostics

	Real male	Real female	Sum
Morphological male	63	4	67
Morphological female	2	29	31
	65	33	98

Both sexes can be estimated with about 94% accuracy. Sensitivity=0.97, Specificity=0.88; n=98.

methods [32, 33, 40, 47, 48]. Furthermore, Kappa statistics were applied to measure or quantify the disagreement between sex diagnoses done by different observers [2, 3, 40]. For this reason, all digitally collected measurements from the two examiners were entered into the software solution Fordisc® in order to test differences in the sex classification resulting from measurement fluctuations. Finally, kappa statistics were calculated for morphological sexing by using the sum scores from the morphological examination done by two different observers for each case.

Results

Of all 98 crania, 67 were assessed to be male and 31 female using the nonmetric traits on the skeletonized skull (Table 4). As shown in Table 2, the sex of 92 crania was correctly classified by using morphological methods. In contrast, of the 98 crania; Fordisc® 3.0 classified only 84 skulls in the correct group (χ^2=3.56, two-sided p value=0.06). Yates' correction for continuity (which adjusts Pearson's χ^2 test) was not necessary in this sample.

The probability that Fordisc® will indicate a male cranium when the particular cranium was proven male, i.e., by autopsy or DNA analysis, was about 0.89 (sensitivity for male; true positive rate for males). The probability that Fordisc® will indicate a female cranium when the cranium was in fact female was about 0.79 (specificity for male). Table 4 shows the corresponding results from morphological analysis (sensitivity 0.97; true negative rate for male or specificity=0.88). Table 5 and Table 6 demonstrate the results of Kappa statistics after metric and morphological analysis done by two different observers, with a value of κ=0.79 for the morphological test and a value of κ=0.93 for Fordisc® [95% CI κ (Fordisc®)=0.85 to 1.00; κ (morph)=0.66 to 0.92]. The area under the ROC curve (AUC) is presented as an overall measure of diagnostic accuracy (Fig. 2). The value for the AUC was 0.92 for metric analysis using Fordisc® (95% confidence interval=0.851 to 0.967) and 0.96 for morphological analysis (95% confidence interval=0.898 to 0.988). No significant differences between the areas under the receiver operating curve (AUC_{diff}=0.036) were observed among the two compared methods (fs$p$=0.14).

Table 5 Kappa statistics: to quantify the disagreement between sex diagnoses given by different observers after discriminant function analysis with Fordisc®

	Expert A: male	Expert A: female	Sum
Expert B: male	63	1	64
Expert B: female	2	32	34

κ=0.93. (95% CI=0.86–1.00). Percentage agreement=96.9%.

Table 6 Kappa statistics: to quantify the disagreement between sex diagnoses given by different observers after morphological analysis

	Expert A: male	Expert A: female	Sum
Expert B: male	61	5	66
Expert B: female	4	28	32

κ=0.79 (95% CI=0.66–0.92). Percentage agreement=90.8%.

Discussion

Previous studies investigating the suitability of discriminant function analysis outside the reference population group for which the metric functions have been developed have shown that the true classification rate is poorer than for traditional morphological methods [26, 41, 43]. We used the χ^2 nonparametric test of statistical significance for bivariate tabular analysis to test the null hypothesis that the classification levels of the two methods have the same frequency. Our results indicated that the classification rates

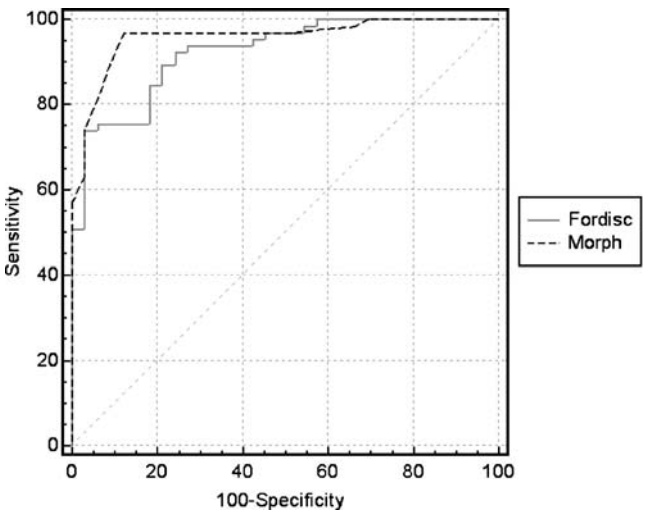

Fig. 2 ROC analysis results: AUC (Fordisc®)=0.92; AUC (Morphol)=0.96; difference between areas=0.036; significance level, p=0.139 (probability of the hypothesis that the difference between the two area under curve is 0)

of the two sexing methods do not differ significantly at a level of α=0.05. The value of p=0.06 should, however, not be misinterpreted as a positive proof for equality of the methods and therefore as an exclusion of any meaningful influence of the reference samples on Fordisc® estimations. Choosing an alternative significance level of α=0.10 leads to significant differences between classification rates. The ability of a test method (software vs morphological method) to discriminate male against female cases was evaluated using ROC. Accuracy was measured by the AUC. When the method cannot distinguish between the two sexes, the AUC will equal 0.5 (the ROC curve will coincide with the diagonal). The results show that there is a good but not perfect separation of the two sex groups, and the AUC does not equal 1. On our skull material, the morphological method tests sex more accurately (AUC=0.96) than Fordisc® (AUC=0.92). These differences are, however, not statistically significant (p=0.14).

Clearly, regardless of how sensitive and specific the test may be, if the result cannot be replicated, the method is of little use. The low rate of personal disagreement among observers (κ_{morph}=0.79; $\kappa_{\text{Fordisc®}}$=0.93) about classification shows, furthermore, that the morphological and the metric methods entail a low level of interobserver risk of error. In general, κ values between 0.61 and 0.8 could be interpreted as good, κ values greater than 0.81 as very good agreement [3]. The sample displayed a typical composition concerning the age and sex profile in forensic cases in Germany and was, therefore, not represented in the database used in Fordisc®. In our sample, derived from forensic cases between 1968 and 2005 with a slight predominance of older cases, both sexes could be estimated with about 86% accuracy using Fordisc® (vs 94% using morphognostics), and males were identified slightly more accurately than females. Because of a different population profile and the secular changes in the past 25 years, this accuracy may be reduced in more recent and less homogeneous forensic samples. Concerning the question of the usability of the software, which forms discriminant functions based on recent US American population databases (FDA), our results seem to make apparent that the use of Fordisc® as a computerized sex estimation tool for forensic cases in Europe cannot be recommended without a built-in software update based on European reference samples. Despite the moderately good classification rates on the skull material in this study, as soon as forensic cases differ from the data base reference sample, any sex classification done automatically by the software Fordisc® has to be interpreted with maximum caution unless the examiner has the opportunity to compare the computer results with morphological observation results. In fact, in forensic cases, it is often difficult or even impossible to determine the membership to a specific population group from the findings or discovery circumstances, and the choice of a reference population for the upcoming analyses therefore remains, in some respects, a game of chance. This reason alone calls for a combined use of both available methods and underlines the necessity of continuous further advancement of the methods as long as DNA-based sex determination does not guarantee proof results for all possible forensic circumstances.

References

1. Acsádi G, Nemeskéri J (1970) History of human life span and mortality. Akadémia Kiadó, Budapest
2. Altman DG (1980) Statistics and ethics in medical research. VI—Presentation of results. Br Med J 281:1542–1544
3. Altman DG (1980) Statistics and ethics in medical research. VII—Interpreting results. Br Med J 281:1612–1614
4. Birkby WH (1966) An evaluation of race and sex identification from cranial measurements. Am J Phys Anthropol 24:21–27
5. Cattaneo C, Grandi M (2004) Antropologia E Odontologia Forense, 1st edn. Monduzzi Editore SpA, Bologna
6. Chang YM, Both K, Burgoyne LA (2002) Genetic variation at three DYS-STR loci and amelogenin between the Australian Caucasians and three Asian ethnic groups. J Forensic Sci 47: 918–921
7. Cox M (2000) Human osteology in archaeology and forensic science. Greenwich Medical Media, London
8. Czarnetzki A (1971) Metrische Merkmale der langen Extremitätenknochen im Populationsvergleich. Homo 31:221–227
9. Di Bennardo R, Taylor JV (1983) Multiple discriminant function analysis of sex and race in the postcranial skeleton. Am J Phys Anthropol 61:305–314
10. Ericksen MF (1982) How "representative" is the Terry collection? Evidence from the proximal femur. Am J Phys Anthropol 59: 345–350
11. Evison MP (1996) HLA: a preliminary survey. Anc Biomol 3:1–28
12. Faerman M, Filon D, Kahila G, Greenblatt CL, Smith P, Oppenheim A (1995) Sex identification of archaeological human remains based on amplification of the X and Y amelogenin alleles. Gene 167:327–332
13. Faerman M, Kahila G, Smith P, Greenblatt C, Stager L, Filon D, Oppenheim A (1997) DNA analysis reveals the sex of infanticide victims. Nature 385:212–213
14. Finnegan M (1978) Non-metric variation of the infracranial skeleton. J Anat 125:23–37
15. Giles E (1964) Sex determination by discriminant function analysis of the mandible. Am J Phys Anthropol 22:129–135
16. Giles E, Elliot O (1963) Sex determination by discriminant function analysis of crania. Am J Phys Anthropol 21:53–68
17. Grubwieser P, Muhlmann R, Berger B, Niederstatter H, Pavlic M, Parson W (2006) A new "miniSTR-multiplex" displaying reduced amplicon lengths for the analysis of degraded DNA. Int J Leg Med 120:115–120
18. Howells WW (1973) Cranial variation in man a study by multivariate analysis of patterns of difference among recent human populations. Papers of the Peabody Museum of Archaeology and Ethnology, Harvard University, Boston
19. Hunt DR, Albanese J (2005) History and demographic composition of the Robert J. Terry anatomical collection. Am J Phys Anthropol 127:406–417

20. Jantz RL (1992) Modification of the Trotter and Gleser female stature estimation formulae. J Forensic Sci 37:1230–1235
21. Jantz RL, Meadows Jantz L (2000) Secular change in craniofacial morphology. Am J Human Biol 12:327–338
22. Jantz RL, Ousley SD (1993) FORDISC 1.0: Computerized forensic discriminant functions. University of Tennessee, Knoxville
23. Jonke E, Schaefer K, Freudenthaler JW, Prossinger H, Bookstein FL (2003) A cephalometric comparison of skulls from different time periods—the Bronze Age, the 19th century and the present. Coll Antropol 27:789–801
24. Kemkes-Grottenthaler A (2001) The reliability of forensic osteology—a case in point. Forensic Sci Int 117:65–72
25. Kemkes-Grottenthaler A (2005) Sex determination by discriminant analysis: an evaluation of the reliability of patella measurements. Forensic Sci Int 147:129–133
26. L'Engle FW, Belcher R, Armlagos GJ (2005) Forensic misclassification of ancient Nubian crania: implications for assumptions about human variation. Curr Anthropol 46:340–346
27. Lassen C, Hummel S, Herrmann B (1996) PCR based sex identification of ancient human bones by amplification of X- and Y-chromosomal sequences: a comparison. Anc Biomol 1:25–34
28. Lassen C, Hummel S, Herrmann B (2000) Molecular sex identification of stillborn and neonate individuals ("Traufkinder") from the burial site Aegerten. Anthropol Anz 58:1–8
29. Martin R (1928) Lehrbuch der Anthropologie, 2nd edn. Fischer, Jena
30. Meadows L, Jantz RL (1995) Allometric secular change in the long bones from the 1800s to the present. J Forensic Sci 40:762–767
31. Meindl RS, Russell KF, Lovejoy CO (1990) Reliability of age at death in the Hamann-Todd collection: validity of subselection procedures used in blind tests of the summary age technique. Am J Phys Anthropol 83:349–357
32. Metz CE (1978) Basic principles of ROC analysis. Semin Nucl Med 8:283–298
33. Metz CE, Herman BA, Roe CA (1998) Statistical comparison of two ROC-curve estimates obtained from partially-paired datasets. Med Decis Mak 18:110–121
34. Moore-Jansen PH, Jantz RL, Ousley SD (1994) Data collection procedures for forensic skeletal material. Forensic anthropology series, 3rd edn. Department of Anthropology, University of Tennessee, Knoxville
35. Nemeskéri J (1970) 1966 évben egyetemi (főiskolai) felvételre jelentkezettek demográfiai és testfejlettségi vizsgálata. Statisztikai Kiadó Vállalat, Budapest
36. Ousley SD, Jantz L (1998) The forensic data bank: documenting skeletal trends in the United States. In: Reichs KJ (ed) Forensic osteology. 2nd edn. Charles C. Thomas, Springfield, pp 441–458
37. Ousley SD, Jantz LM (1996) FORDISC 2.0: Personal computer forensic discriminant functions. University of Tennessee, Knoxville
38. Rösing W, Graw M, Marré B, Ritz-Timme S, Rothschild MA, Rötzscher K, Schmeling A, Schröder I, Geserick G (2004) Forensische Geschlechts—und Altersdiagnose am Skelett. Empfehlungen. Kriminalistik 11:712–716
39. Ross AH, Ubelaker DH, Falsetti AB (2002) Craniometric variation in the Americas. Hum Biol 74:807–818
40. Schoonjans F, Zalata A, Depuydt CE, Comhaire FH (1995) MedCalc: a new computer program for medical statistics. Comput Methods Programs Biomed 48:257–262
41. Sejrsen B, Lynnerup N, Hejmadi M (2005) An historical skull collection and its use in forensic odontology and anthropology. J Forensic Odonto-stomatol 23:40–44
42. Snow CC, Hartman S, Giles E, Young FA (1979) Sex and race determination of crania by calipers and computer: a test of the Giles and Elliot discriminant functions in 52 forensic science cases. J Forensic Sci 24:448–460
43. Ubelaker DH, Ross AH, Graver S (2002) Application of forensic discriminant functions to a Spanish cranial sample. Forensic Sci Commun 4:
44. von Wurmb-Schwark N, Bosinski H, Ritz-Timme S (2007) What do the X and Y chromosomes tell us about sex and gender in forensic case analysis? J Clin Forensic Med 14:27–30
45. Wiegand P, Klein R, Braunschweiger G, Hohoff C, Brinkmann B (2006) Short amplicon STR multiplex for stain typing. Int J Leg Med 120:160–164
46. Işcan YM, Steyn M (1999) Craniometric determination of population affinity in South Africans. Int J Leg Med 112:91–97
47. Zweig MH (1993) ROC plots display test accuracy, but are still limited by the study design. Clin Chem 39:1345–1346
48. Zweig MH, Campbell G (1993) Receiver-operating characteristic (ROC) plots: a fundamental evaluation tool in clinical medicine. Clin Chem 39:561–577