Received: 10 September 2023 | Revised: 31 December 2023 | Accepted: 3 May 2024
DOI: 10.1002/oa.3308

RESEARCH ARTICLE

Nonmetric population-specific sex estimation based on the skull using logistic regression for Flemish samples

Maggie Wongsantativanich | Isabelle De Groote

Department of Archaeology, ArcheOs: Research Laboratory for the study of Biological Anthropology, Ghent University, Ghent, Belgium

Correspondence
Maggie Wongsantativanich, Department of Archaeology, ArcheOs: Research Laboratory for the study of Biological Anthropology, Ghent University, Sint-Pietersnieuwstraat 35, 9000 Ghent, Belgium.
Email: maggiegaudencio@gmail.com

Abstract

There are very few sex estimation methods specifically designed for or tested on Belgian skulls. The currently used methods for European populations have been developed using North American collections where individuals are categorized as White and/or having European ancestry. These frequently show discordance between the pelvic sex and cranial sex estimations highlighting the need for population specific methods. To fill this gap in our knowledge, several sex estimation methods, using 15 qualitative skull features, were tested on two Flemish (northern Belgium) skeletal collections; one archaeological (15th–17th century) and one forensic (20th century). The features were tested by themselves as well as in different combinations using logistic regression. The glabella is considered the best lone feature with a minimal accuracy of 78.4% and a sex bias of −5.2%. Furthermore, four sex estimation equations were developed for the skull, the cranium, the mandible, and the frontal bone separately. The skull has an accuracy of 89.3% and a bias of 0.8%. For the cranium, this is 87.5% and −0.3%, respectively, for the mandible 85.1% and −0.1%, and for the frontal bone it is 80.4% and −4.6%. The various tests confirm that many skull features can be used for sex estimation and can generate high sex estimation accuracy.

KEYWORDS
archaeology, Belgium, biometric profiling, qualitative, region-specific, sex determination

1 | INTRODUCTION

Archaeologists seek to understand humanity's past through the study of material remains, human remains included. The study of human remains allows insight into the health and demographics of past populations as well as some cultural practices. However, these studies do require ethical considerations as they represent people and not solely objects, and to meaningfully study health, lifeways, and demography of past populations, the biological profile of individuals must be determined (Buikstra & Ubelaker, 1994; White et al., 2012; White & Folkens, 2005). Sex estimation is therefore one of the most important factors in the archaeological study, and it influences the determination of other biological markers such as age and stature.

It is to be noted that within the context of this study, the term "sex," as well as "male" and "female," is used to define biological sex, and not gender (which is a social and cultural construct) (Christensen et al., 2014; Mays & Cox, 2000; White et al., 2012).

There are very few nonmetric sex estimation methods specifically designed for or tested on European populations (Boucherie et al., 2021; Capitaneanu et al., 2017). Also, many popular nonmetric methods that are used for European skeletons were developed using

North American collections, where individuals are categorized as White, meaning they have European ancestry, but are not more specific (Garvin et al., 2014; Rogers, 2005; Walker, 2008; Williams & Rogers, 2006). Given the genetic and morphological variation between populations due to the isolation-by-distance model (Relethford, 2004), which includes differences in sexual dimorphism (Pickering & Bachman, 2009; White & Folkens, 2005), it is useful to test known methods on new skeletal collections and attempt to develop new methods for specific populations. The need for such estimation methods on Flemish skulls was identified during research on the collections (De Groote et al., 2023) where it was often found that the pelvis and cranium were inconsistent in their sex estimation when standard methods were used (Buikstra & Ubelaker, 1994). De Groote and Humphrey (2016) also observed the need for population-specific sex estimation methods when studying the Iberomaurusian Later Stone Age hunter–gatherers, where all individuals scored on the masculine scale of the cranial traits. Bruzek et al. (2004) confirmed this discrepancy between the pelvis and skull sex indicators for these same individuals. A recent study by Kamnikar et al. (2022) investigated regional differences within the "Hispanic" ancestry population of Columbia. They observed small regional differences in cranial morphology reflecting regional genetic groups, again showing that differences between groups exist beyond the ancestry classification, and it should be considered that these may also affect sex estimation methods. Anthropologists studying diverse human osteological collections are best placed to estimate the need for region-specific sex-estimation methods.

Several metric sex estimation methods using discriminant functions exist for various European populations such as the Finns (Kajanoja, 1966), Portuguese (Pons, 1955), Croatians (Šlaus & Tomičić, 2005), and Greeks (Kiskira et al., 2022), but none of these share the population history of Belgium.

The Belgian population can genetically be distinguished from other European (i.e., Finnish, British, Spanish, and Italian) populations, and there are even subtle differences between Flanders and Wallonia (Van den Eynden et al., 2018). Therefore, in this study, various skull features are tested for their use in sex estimation methods on a Flemish sample.

2 | MATERIALS AND METHODS

2.1 | Materials

The total study sample consists of 178 skulls coming from three separate collections.

Ten skeletons from an archaeological excavation of the Sint-Pietersplein in Ghent (12th–13th century) (Boudin et al., 2009) were used exclusively for the intraobserver error test. These were not included in the dataset for the statistical sex estimation analysis. The ten skulls and respective pelvic bones were randomly selected with only their state of preservation and skeletal maturity as a criterion.

For the sex estimation analyses, two collections were used. First is an archaeological one from Aalst (East-Flanders, Belgium). These skeletons were found in 2016 during excavations in the cemeteries of an old monastery. The graves date from the 15th to 17th centuries (Bruggeman et al., 2019, 2021). The actual sex of these individuals is unknown. Therefore, an estimate had to be made using the pelvic bones, because these are considered most accurate for sex determination (Buikstra & Ubelaker, 1994; Klales, 2020). The three Phenice traits (ventral arc, subpubic concavity, and ischiopubic ramus) (Phenice, 1969), the greater sciatic notch, preauricular sulcus (Buikstra & Ubelaker, 1994), subpubic angle, shape of the sacrum, curve of the sacrum, and shape of the obturator foramen (Buikstra & Mielke, 1985; Burns, 2012; Christensen et al., 2014; Pickering & Bachman, 2009) were used to estimate sex. Only 80 (31 male, 49 female) of the 93 skeletons were used for this study, as highly fragmented, pathological, and juvenile skeletons were excluded.

The second collection is a forensic one of known sex curated by the Ghent University Museum. The full collection includes 222 human skulls. The individuals died between the 1930s and 1990s, and given the forensic nature of the collection, all had legal case numbers and files attached to them. For this study, the only personal information that was obtained was the following: sex, age-at-death (year of birth and death), and place of birth. Children, individuals of unknown sex and/or age, and individuals not born in Belgium were eliminated. The collection contained a majority of male skulls, and thus only 38 suitable female skulls could be selected. In the end, 88 skulls were found suitable for this study, of which 38 were female and 50 were male.

2.2 | Method: selected skull features

Fifteen morphological skull features/traits were selected for study. The use of these features is largely based on the studies of Buikstra and Ubelaker (1994), Rogers (2005), Williams and Rogers (2006), and Walker (2008). The majority of the selected traits exhibit good intraobserver agreement (Garvin et al., 2014; Walker, 2008) and in some cases also good interobserver agreement (Walker, 2008; Walrath et al., 2004). Some of these features are scored to reflect the expression of a particular trait (e.g., 1 to 5), while others require the observer to decide on whether the feature is expressed as female or male (M, F, I).

(1) First, the *glabella* and supraorbital ridges were selected following the example of Buikstra and Ubelaker (1994) (see Figure 1). This feature is easy to view and is recommended in multiple studies (Garvin et al., 2014; Walker, 2008; Williams & Rogers, 2006).

(2) The same goes for the *mastoid process*.

(3) The *external occipital protuberance*, despite scoring poorly in Walker (2008) and Garvin et al. (2014), is also included in this study. It is easy to observe and Walrath et al. (2004) argue that a clear description can improve results. In the study of Buikstra and Ubelaker (1994), however, the illustration and description contradict each other.

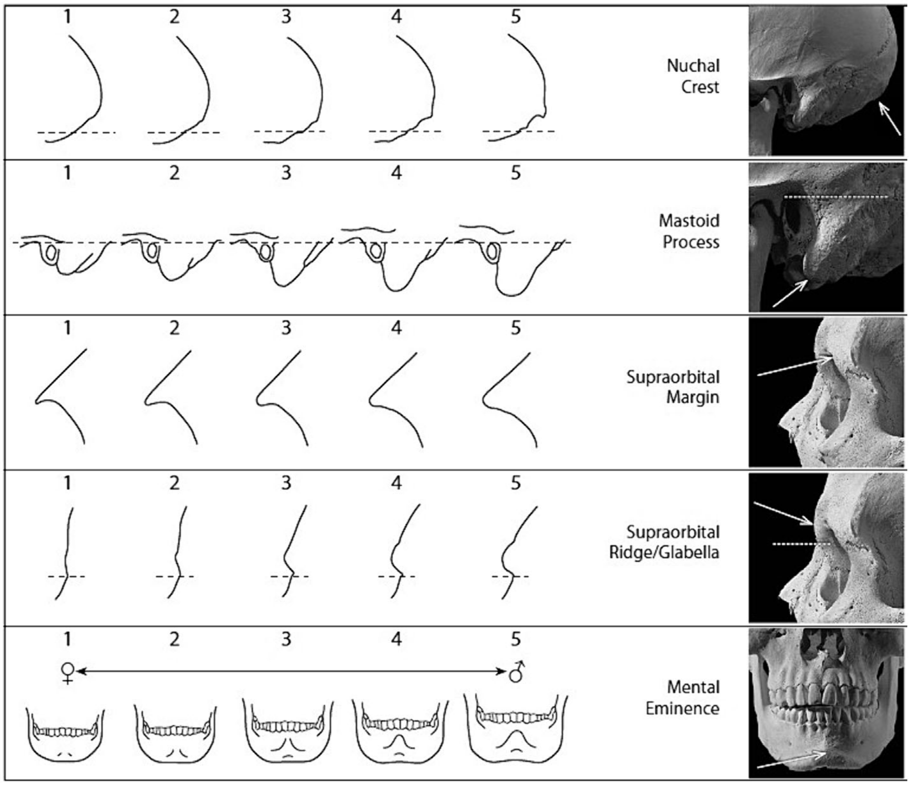

FIGURE 1 Scoring system for the nuchal crest, mastoid process, supraorbital margin, glabella, and mental eminence by Walker (image from White et al., 2012).

Therefore, it was decided to divide the nuchal crest into two categories: the actual external occipital protuberance or inion hook and then the nuchal markings. Here the external occipital protuberance refers to the appearance of the occipital bone in *norma lateralis*. For this, we look at Figure 1.

(4) The *nuchal markings* refer to the surface of the posterior part of the occipital. This involves looking at the nuchal lines and general rugosity of the region. This is then considered gracile/female (F), rugose/male (M), or intermediate (I).

(5) (6) The *supraorbital margin* was also observed in two ways. First, there is a "tactile" approach, following Buikstra and Ubelaker (1994) (see Figure 1). The second approach, following Graw et al. (1999) (see Figure 2) involves taking a piece of polymer clay and taking an impression of the supraorbital ridge. The tactile method scored poorly with Walker (2008), but the *imprint* method possibly offers a more objective and accurate approach.

(7) The *nasal aperture* is included in this study because it was the best lone sex indicator in Rogers (2005) and was again recommended in Williams and Rogers (2006). This characteristic is rarely mentioned in other literature and should therefore be studied more. The size, placement and sharpness of the lower border of the nasal aperture is observed (following Rogers' (2005) example). In males, the nasal aperture is said to be longer, appears to be positioned higher on the face and the edge is sharp. In females it is wider, lower and the edge is rounded (Rogers, 2005). In this study this trait is marked as female, male, or intermediate.

(8) The *zygomatic extension* refers to the posterior root of the zygomatic bone. In males, the root allegedly extends beyond the external auditory meatus into the supramastoid crest, which in turn becomes part of the temporal line (Burns, 2012; Rogers, 2005). In females, the root ends at the external auditory meatus (Burns, 2012).

(9) Furthermore, it was also decided to register the shape of the *marginal tubercle* on the frontal process of the zygomatic bone to see if there is a connection to sex. During the intraobserver error test observations, MW noted that the marginal tubercle differed between a male and a female individual, where the male had a much more prominent tubercle than the female. MW, therefore, wanted to test whether this was indeed a sexually dimorphic feature that can be used as a sex indicator. Silva et al. (2020), in a study on Chilean skulls, found that the tubercle was indeed more prominent in males than females.

(10) Finally, the size and general *rugosity* of the skull are included in this study, following the example of Williams and Rogers (2006).

In addition to the above-mentioned cranial features, several mandibular features were also included.

(11) The *mental eminence* was scored based on the 1 to 5 scoring system in Figure 1. This characteristic was the worst indicator in the study by Garvin et al. (2014), but is nevertheless included to use all five traits used by Buikstra and Ubelaker (1994).

(12) (13) (14) In addition, the *shape of the chin*, the prominence of *gonial eversion*, and the *gonial angle* were examined, as these are practically useful morphological features of the mandible. The gonial angle was also recommended by Williams and Rogers (2006).

(15) Finally, the *mandibular ramus flexure* was also included. This is a flexure on the ramus at the level of the occlusal surface of the molars, which according to Loth and Henneberg (1996) only occurs in males (see Figure 3).

The chin shape, gonial eversion, gonial angle, and ramus flexure were again 'scored' as male, female, or intermediate.

For more details on how to score the different skull traits, see Appendix S1.

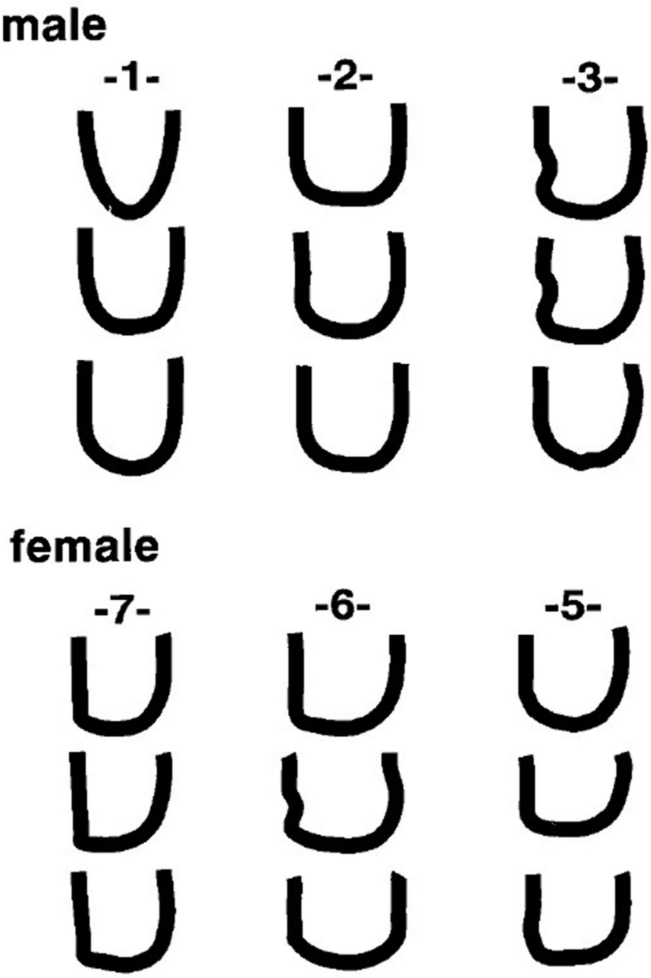

FIGURE 2 Scoring system for the imprint of the supraorbital margin by Graw et al. (1999).

2.3 | Methods: statistics

Several statistical methods were used to test the relationship between the given feature/trait scores and sex. Trait scores scored as female (F), intermediate (I), or male (M) in the dataset were transformed in the statistical dataset to 1, 2, and 3, respectively, to obtain ordinal variables. Missing values were replaced by the medians of the males and females in multivariate analyses, to obtain a complete dataset. Only the nasal aperture values were not replaced in the archaeological sample and therefore omitted, because there were more missing values than valid ones, so the medians were not considered reliable.

First, an intraobserver error test was performed, using the aforementioned 10 skulls from Ghent, to assess whether one researcher can score the different skulls traits consistently over time. The skulls were assessed on three separate occasions by the same person over the course of 3 weeks. The scores given to the skulls on the three different occasions were then compared with each other. The rates of perfect agreement (i.e., the given scores were exactly the same for all observations) and of approximate agreement (i.e., the given score differed, but by no more than 1 digit) were determined.

A Mann–Whitney *U* test (using JASP, version 0.14) was then performed to see whether there is a statistical difference between the trait scores given to males and females. This test was selected because the data are semiquantitative so a nonparametric test was needed. The samples were considered independently as well as pooled.

Furthermore, logistic regressions (using SPSS, version 27.0) were performed to determine which (combination of) features can be used to estimate the sex of an individual. Walker (2008) gathered that logistic regression models provided the best results for ordinal skull feature scores, as they had low misclassification rates and a small sex bias. In this study, both standard and stepwise logistic regressions were used and compared. First a series of standard logistic regressions were performed using different sets of skull features (e.g., all features, only the cranial features, only the mandibular features, and only the frontal features). This was done for both the archaeological and forensic samples individually, as well as pooled. Then a series of stepwise regressions was done as well in the same fashion. Afterwards, the regressions based on the archaeological sample were tested on the forensic sample and vice versa. The various logistic regressions were then compared with one another, and the

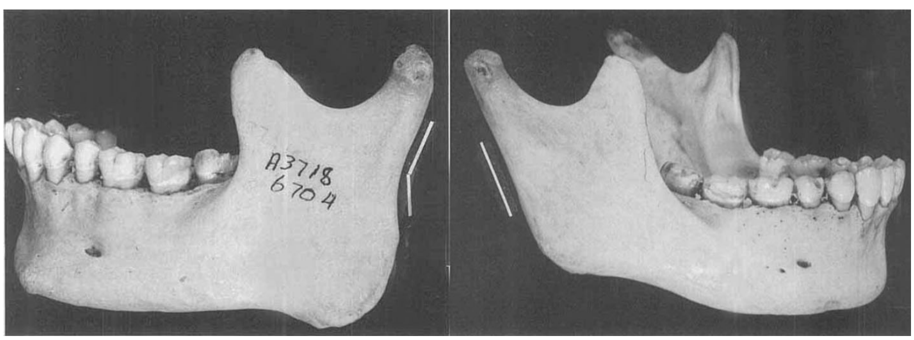

FIGURE 3 (Left) an adult male mandible displaying a mandibular ramus flexure. (Right) an adult female mandible without a mandibular ramus flexure (photo by M. Henneberg, from Loth & Henneberg, 1996).

best methods were selected and highlighted. This selection of best methods was based on the following criteria of both the archaeological and forensic (and thus also the pooled) samples:

- the percentage of correctly classified individuals (≥80%)
- and the sex bias (≤|10|%) (= % correctly classified females − % correctly classified males; this number shows whether more males (positive number) or more females (negative number) are correctly classified, ideally the sex bias would be 0%).

Then, from the results of the logistic regressions, discriminatory functions were derived that can be used to estimate the sex of a skull.

Due to the many logistic regressions being discussed in this work, all regressions are given a "code" to make referencing easier. The regressions performed on the archaeological sample are all given the letter "A," the forensic sample is given letter "F," and the grouped (or pooled) sample gets letter "P." These then get one or several numbers or Roman numerals added on. The numbers denote a standard logistic regression and the Roman numerals denote a stepwise logistic regression.

3 | RESULTS

3.1 | Intraobserver error test

For an in-depth discussion of the intraobserver error test, see Appendix S1.

At first, the rates of perfect intraobserver agreement were overall rather low, between 20% and 86% (see Table S1). The degree of approximate agreement was between 50% and 100% (see Table S2). Therefore, several extra guidelines were established to make the scoring of traits as consistent as possible (see Appendix S1). This resulted in higher rates of intraobserver agreement in the end, resulting in approximate agreement rates between 78% and 100%.

The pelvic traits, used for sex determination of the archaeological sample, had a high rate of perfect intraobserver agreement (89% to 100%) and did not need further refinement.

3.2 | Difference between male and female trait scores

A Mann–Whitney *U* test was performed to compare females to males.

The archaeological sample expressed a significant difference between males and females for almost all traits (see Table S3). The only traits that showed no significant difference were the supraorbital margin (tactile) (median males = 3, median females = 2, $U = 517.5$, $p = 0.054$), zygomatic extension (median males = 2, median females = 1, $U = 498$, $p = 0.064$), and marginal tubercle (median males and females = 2, $U = 523$, $p = 0.25$). In the forensic sample (see Table S4), there was a significant difference between males and females for all traits. When the two samples are pooled a significant difference is again detected between males and females for each skull trait (see Table S5).

3.3 | Logistic regression

3.3.1 | Univariate models

A series of univariate logistic regressions were performed based on all the individual skull features by themselves. This was done for the archaeological, forensic, and pooled samples (see Tables 1–3).

In the archaeological sample, sex was correctly classified for 65% to 83% of the individuals, depending on the used trait (Table 1); 60% to 81% of individuals were correctly classified in the forensic sample (Table 2). The logistic regression model for the pooled sample was significant for all skull features with 60% to 79% of skulls classified correctly (Table 3).

Most traits in all three samples showed significant results. However, many traits also had a high sex bias (>|10|%).

3.3.2 | Multivariate models

Archaeological sample

When all traits are entered in a standard multivariate logistic regression (A.1), a significant model ($\chi^2(40) = 106.819$, $p < 0.001$) with 100% correct classification is obtained (Table 4). When traits that were not significant by themselves (i.e., mastoid process, supraorbital margin tactile, zygomatic extension, and marginal tubercle; see Section 3.3.1 and Table 1) are removed, the model is still significant ($\chi^2(28) = 106.819$, $p < 0.001$) with a correct classification of 100%. If traits with a high sex bias are also removed (i.e., external occipital protuberance, supraorbital margin imprint, and chin shape), the model (A.1.1) still obtains 100% correct classification ($\chi^2(17) = 106.819$, $p < 0.001$).

Because in the context of archaeological excavations the full skull is sometimes not found, a model containing only the cranial features (without the mandible) was tested too. When all cranial features are used (except nasal aperture), the model (A.2) is statistically significant ($\chi^2(28) = 104.046$, $p < 0.001$) with a high correct classification of 98.8% and a sex bias of −2%.

When only the mandible features are used (A.3) ($\chi^2(12) = 66.767$, $p < 0.001$), 87.5% of skulls are correctly classified with a sex bias of 5.9%.

To test whether individual bones, like the temporal bone, could be used for sex estimation, a model using only the mastoid process and zygomatic extension was performed as well. This model (A.4) ($\chi^2(6) = 16.657$, $p = 0.011$) correctly classified 70% of skulls; however, the sex bias was 30%, indicating most skulls were classified as female.

The frontal bone fared better. The best combination for the frontal bone used the supraorbital margin (tactile) and glabella. This method (A.5) ($\chi^2(7) = 32.834$, $p < 0.001$) resulted in a correct classification of 80% and a bias of −6.3%.

TABLE 1 Results of the univariate logistic regressions on the various skull features in the archeological sample. Significant p values ($p < 0.05$), total percentages ≥75%, and gender biases ≤|10|% are in bold. Sex bias = % correctly classified females − % correctly classified males.

Trait archeological sample, $n = 80$	p value	Observed		Predicted Sex: Female	Predicted Sex: Male	Correct (%)	Sex bias (%)
Glabella ($n = 67$)	**<0.001**	Sex	Female	32	11	74.4	**−4.7**
			Male	5	19	79.2	
		Total percentage				**76.1**	
Mastoid process ($n = 79$)	0.073	Sex	Female	42	7	85.7	45.7
			Male	18	12	40	
		Total percentage				68.4	
External occipital protuberance ($n = 70$)	**<0.001**	Sex	Female	44	0	100	53.8
			Male	14	12	46.2	
		Total percentage				**80**	
Nuchal markings ($n = 75$)	**<0.001**	Sex	Female	35	12	74.5	**−4.1**
			Male	6	22	78.6	
		Total percentage				**76**	
Supraorbital margin (tactile) ($n = 75$)	0.37	Sex	Female	41	6	87.2	55.1
			Male	19	9	32.1	
		Total percentage				66.7	
Supraorbital margin (imprint) ($n = 75$)	**<0.01**	Sex	Female	43	4	91.5	45.1
			Male	15	13	46.4	
		Total percentage				74.7	
Nasal aperture ($n = 22$)	0.071	Sex	Female	14	1	93.3	50.4
			Male	4	3	42.9	
		Total percentage				**77.3**	
Zygomatic extension ($n = 74$)	0.153	Sex	Female	49	0	100	96
			Male	24	1	4	
		Total percentage				67.6	
Marginal tubercle ($n = 71$)	0.502	Sex	Female	40	6	87	63
			Male	19	6	24	
		Total percentage				64.8	
Rugosity ($n = 70$)	**<0.001**	Sex	Female	33	12	73.3	−10.7
			Male	4	21	84	
		Total percentage				**77.1**	
Mental eminence ($n = 74$)	**<0.001**	Sex	Female	36	9	80	14.5
			Male	10	19	65.5	
		Total percentage				74.3	
Chin shape ($n = 74$)	**<0.001**	Sex	Female	41	4	91.1	35.9
			Male	13	16	55.2	
		Total percentage				**77**	
Mandibular ramus flexure ($n = 77$)	**<0.001**	Sex	Female	37	10	78.7	**5.4**
			Male	8	22	73.3	
		Total percentage				**76.6**	
Gonial eversion ($n = 78$)	**<0.001**	Sex	Female	36	12	75	11.7
			Male	11	19	63.3	
		Total percentage				70.5	
Gonial angle ($n = 77$)	**<0.001**	Sex	Female	42	5	89.4	16.1
			Male	8	22	73.3	
		Total percentage				**83.1**	

TABLE 2 Results of the univariate logistic regressions on the various skull features in the forensic sample. Significant *p* values ($p < 0.05$), total percentages ≥75%, and gender biases ≤|10|% are in bold. Sex bias = % correctly classified females − % correctly classified males.

Trait forensic sample, $n = 88$	*p* value	Observed		Predicted Sex: Female	Predicted Sex: Male	Correct (%)	Sex bias (%)
Glabella ($n = 86$)	**<0.001**	Sex	Female	28	8	77.8	−**4.2**
			Male	9	41	82.0	
		Total percentage				**80.2**	
Mastoid process ($n = 88$)	**<0.01**	Sex	Female	22	16	57.9	−10.1
			Male	16	34	68.0	
		Total percentage				63.6	
External occipital protuberance ($n = 97$)	**<0.001**	Sex	Female	20	18	52.6	−31.1
			Male	8	41	83.7	
		Total percentage				70.1	
Nuchal markings ($n = 87$)	**<0.001**	Sex	Female	17	21	44.7	−45.1
			Male	5	44	89.8	
		Total percentage				70.1	
Supraorbital margin (tactile) ($n = 88$)	**<0.001**	Sex	Female	27	11	71.1	−**2.9**
			Male	13	37	74	
		Total percentage				72.7	
Supraorbital margin (imprint) ($n = 88$)	0.068	Sex	Female	14	24	36.8	−47.2
			Male	8	42	84	
		Total percentage				63.6	
Nasal aperture ($n = 81$)	**0.046**	Sex	Female	27	7	79.4	32.6
			Male	25	22	46.8	
		Total percentage				60.5	
Zygomatic extension ($n = 88$)	**0.03**	Sex	Female	30	8	78.9	30.9
			Male	26	24	48	
		Total percentage				61.4	
Marginal tubercle ($n = 88$)	0.141	Sex	Female	16	22	42.1	−31.9
			Male	13	37	74	
		Total percentage				60.2	
Rugosity ($n = 88$)	**<0.001**	Sex	Female	27	11	71.1	−16.9
			Male	6	44	88	
		Total percentage				**80.7**	
Mental eminence ($n = 87$)	**<0.01**	Sex	Female	16	21	43.2	−46.8
			Male	5	45	90	
		Total percentage				70.1	
Chin shape ($n = 87$)	**<0.001**	Sex	Female	25	12	67.6	−12.4
			Male	10	40	80	
		Total percentage				74.7	
Mandibular ramus flexure ($n = 88$)	**<0.01**	Sex	Female	28	10	73.7	11.7
			Male	19	31	62	
		Total percentage				67	
Gonial eversion ($n = 88$)	**<0.001**	Sex	Female	20	18	52.6	−31.4
			Male	8	42	84	
		Total percentage				70.5	
Gonial angle ($n = 88$)	**<0.001**	Sex	Female	20	18	52.6	−23.4
			Male	12	38	76	
		Total percentage				65.9	

TABLE 3 Results of the univariate logistic regressions on the various skull features in the pooled sample. Significant p values ($p < 0.05$), total percentages ≥75%, and gender biases ≤|10|% are in bold. Sex bias = % correctly classified females − % correctly classified males.

				Predicted			
				Sex			
Trait pooled sample, *n* = 168	***p* value**	**Observed**		**Female**	**Male**	**Correct (%)**	**Sex bias (%)**
Glabella (*n* = 153)	**<0.001**	Sex	Female	60	19	75.9	−**5.2**
			Male	14	60	81.1	
		Total percentage				**78.4**	
Mastoid process (*n* = 167)	**<0.001**	Sex	Female	73	14	83.9	38.9
			Male	44	36	45	
		Total percentage				65.3	
External occipital protuberance (*n* = 157)	**<0.001**	Sex	Female	72	10	87.8	27.8
			Male	30	45	60	
		Total percentage				74.5	
Nuchal markings (*n* = 162)	**<0.001**	Sex	Female	52	33	61.2	−24.5
			Male	11	66	85.7	
		Total percentage				72.8	
Supraorbital margin (tactile) (*n* = 163)	**<0.01**	Sex	Female	51	34	60	−**9.2**
			Male	24	54	69.2	
		Total percentage				64.4	
Supraorbital margin (imprint) (*n* = 163)	**<0.01**	Sex	Female	70	15	82.4	32.4
			Male	39	39	50	
		Total percentage				66.9	
Nasal aperture (*n* = 103)	**<0.01**	Sex	Female	41	8	83.7	37.4
			Male	29	25	46.3	
		Total percentage				64.1	
Zygomatic extension (*n* = 162)	**0.016**	Sex	Female	62	25	71.3	22
			Male	38	37	49.3	
		Total percentage				61.1	
Marginal tubercle (*n* = 159)	**0.029**	Sex	Female	69	15	82.1	46.1
			Male	48	27	36	
		Total percentage				60.4	
Rugosity (*n* = 158)	**<0.001**	Sex	Female	60	23	72.3	−14.4
			Male	10	65	86.7	
		Total percentage				**79.1**	
Mental eminence (*n* = 161)	**<0.001**	Sex	Female	64	18	78	17.2
			Male	31	48	60.8	
		Total percentage				69.6	
Chin shape (*n* = 161)	**<0.001**	Sex	Female	56	26	68.3	−12.7
			Male	15	64	81	
		Total percentage				74.5	
Mandibular ramus flexure (*n* = 165)	**<0.001**	Sex	Female	67	18	78.8	15.1
			Male	29	51	63.7	
		Total percentage				71.5	
Gonial eversion (*n* = 166)	**<0.001**	Sex	Female	40	46	46.5	−41
			Male	10	70	87.5	
		Total percentage				66.3	
Gonial angle (*n* = 165)	**<0.001**	Sex	Female	75	10	88.2	30.7
			Male	34	46	57.5	
		Total percentage				73.3	

TABLE 4 Performance of the different sex determination methods on the different samples. The sample on which the method is based is in italic. Methods with a correct classification rate ≥80% and a sex bias ≤|10|% are in bold. Methods that work well for all samples are underlined. Sex bias = % correctly classified females − % correctly classified males.

Method		Archeological		Forensic		Pooled	
		Correctly classified (%)	Sex bias (%)	Correctly classified (%)	Sex bias (%)	Correctly classified (%)	Sex bias (%)
Skull	A.1	***100***	***0***	72.7	10.9	85.7	10.6
	A.1.1	***100***	***0***	76.1	28.1	87.5	18.8
Cranium	A.2	***98.8***	***−2***	65.9	18.3	81.5	16.8
Mandible	A.3	***87.5***	***5.9***	68.2	28.2	78	24.2
Temporal	A.4	*70*	*30*	62.5	33.6	66.1	32.2
Frontal	A.5	***80***	***−6.3***	**80.7**	**−3.1**	**80.4**	**−4.6**
Walker	A.6	***91.3***	***6.8***	71.6	−24.1	**81**	**−5.8**
Cranial Walker	A.7	***86.3***	***3.9***	71.6	−24.1	78.6	8
Skull	A.I	*92.5*	*14.1*	70.5	−26.7	**81**	**−1**
Cranium	A.II	*88.8*	*18.5*	79.5	22.1	83.9	21.4
Mandible	A.III	*87.5*	*11.2*	67	30.2	76.8	24.3
Skull	F.1	/	/	***100***	***0***	/	/
	F.1.1	68.8	−24.7	***100***	***0***	85.1	−16.8
Cranium	F.2	/	/	***100***	***0***	/	/
	F.2.1	68.8	12.2	***93.2***	***7.4***	**81.5**	**4.9**
Mandible	F.3	86.3	−11.9	***83***	***−7.1***	**84.5**	**−8.4**
Temporal	F.4	68.8	−8.9	*70.5*	*−8.2*	69.6	−8.6
Frontal	F.5	77.5	15.9	*79.5*	*3.6*	78.6	8.7
Walker	F.6	75	38.2	***87.5***	***3.5***	81.5	16.8
Cranial Walker	F.7	77.5	31.7	***86.4***	***0.8***	82.1	13.2
Skull	F.I	83.8	10.3	*83*	*−11.7*	**93.3**	**−1.2**
Cranium	F.II	**85**	**1.8**	***89.8***	***−0.5***	**87.5**	**−0.3**
Mandible	F.III	83.8	−10.7	***83***	***−2.4***	**83.3**	**−6**
Skull	P.1	**100**	**0**	**100**	**0**	***100***	***0***
	P.1.1	**92.5**	**3.6**	**86.4**	**−3.8**	***89.3***	***0.8***
Cranium	P.2	93.8	10.9	**87.5**	**3.5**	***90.5***	***7.9***
Mandible	P.3	**90**	**−5.8**	**80.7**	**1.6**	***85.1***	***−0.1***
Temporal	P.4	68.8	−8.9	70.5	−8.2	*69.6*	*−8.6*
Frontal	P.5	**80**	**−6.3**	**80.7**	**−3.1**	***80.4***	***−4.5***
Walker	P.6	88.8	18.5	**84.1**	**0.2**	***86.3***	***9.3***
Cranial Walker	P.7	88.8	13.2	**85.2**	**−1.8**	***86.9***	***5.7***
Skull	P.I	**92.5**	**8.8**	**88.6**	**1.5**	***90.5***	***5.4***
Cranium	P.II	**85**	**1.8**	**89.8**	**−0.5**	***87.5***	***−0.3***
Mandible	P.III	**88.8**	**−7.8**	**80.7**	**−3.1**	***84.5***	***−3.6***

Furthermore, Walker's method (Walker, 2008), using the glabella, mastoid process, external occipital protuberance, supraorbital margin (tactile), and mental eminence, was tested too. This resulted in a significant model (A.6) ($\chi^2(18) = 69.725$, $p < 0.001$) with 91.3% correctly classified skulls and a bias of 6.8%.

Finally, the cranial features (i.e., excluding the mental eminence of the mandible) from Walker (2008) were tested separately as well. The model (A.7) was significant ($\chi^2(14) = 57.24$, $p < 0.001$), with 86.3% correct classification and 3.9% sex bias.

Three stepwise logistic regressions were performed as well for the full skull, the cranium, and the mandible. The first model (A.I), using the external occipital protuberance, nuchal markings, and gonial angle, was significant ($\chi^2(7) = 85.652$, $p < 0.001$) with a correct classification rate of 92.5% and a sex bias of 14.1%. For the second model

(A.II) ($\chi^2(11) = 68.933$, $p < 0.001$), using the glabella, supraorbital margin imprint and rugosity, this was 88.8% and 18.5%, respectively. The third model (A.III) ($\chi^2(6) = 61.342$, $p < 0.001$) using the shape of the chin, gonial eversion, and gonial angle correctly determined 87.5% of the skulls with a bias of 11.2%.

While some of the above-mentioned methods seem promising, this only applies to the sample on which the method is based. So, method A.1.1 results in 100% correction classification for the archaeological skulls, but when the same formula is applied to the forensic sample, a correct classification of only 76.1% is obtained (see Table 4, A.1.1). The functions based on the archaeological collection do not seem to work well on the forensic sample. The only exception is method A.5 for the frontal bone, which yields even slightly better results on the forensic sample.

Forensic sample

Like with the archaeological sample, 100% correct classification is obtained when all 15 skulls traits are used (F.1) ($\chi^2(42) = 120.352$, $p < 0.001$) (Table 4). This is also the case when the traits that were not significant on their own (i.e., supraorbital margin imprint and marginal tubercle) and the nasal aperture are removed (F.1.1) ($\chi^2(33) = 120.352$, $p < 0.001$).

When only cranial features are used, 100% correct classification is obtained as well (F.2) ($\chi^2(30) = 120.352$, $p < 0.001$). If the solo insignificant features and the nasal aperture are removed (F.2.1) ($\chi^2(21) = 87{,}351$, $p < 0{,}001$), the correct classification rate is 93.2% with a bias of 7.4%.

Of the forensic skulls, 83% are correctly classified only using the mandible features (F.3) ($\chi^2(12) = 51.079$, $p < 0.001$) with a sex bias of −7.1%. For the temporal bone (F.4) ($\chi^2(6) = 19.312$, $p < 0.01$), sex is estimated correctly 70.5% of the time, with a bias of −8.2%. The use of the frontal bone (F.5) ($\chi^2(7) = 48.015$, $p < 0.001$) results in 79.5% correct classification and 3.6% sex bias. The five Walker traits (F.6) result in a significant model ($\chi^2(19) = 74.07$, $p < 0.001$) with 87.5% accuracy and a bias of 3.5%. Without the mental eminence (F.7) ($\chi^2(15) = 70.69$, $p < 0.001$), this is 86.4% and 0.8%, respectively.

When entering all skull traits in a stepwise logistic regression (F.I), a significant model ($\chi^2(4) = 54.71$, $p < 0.001$) is obtained using rugosity and chin shape. The accuracy is then 83% and the bias −11.7%. If only cranial traits are entered (F.II), we get a significant model ($\chi^2(9) = 69.229$, $p < 0.001$) (using the glabella, external occipital protuberance, and rugosity) with an accuracy of 89.8% and a bias of −0.5%. For the mandibular traits (F.III), a significant model is also obtained ($\chi^2(6) = 43.458$, $p < 0.001$) (using chin shape, mandibular ramus flexure, and gonial eversion) with an accuracy of 83% and a sex bias of −2.4%.

When the methods based on the forensic sample are tested on the archaeological one, the results are often similar or sometimes even better (see Table 4). For example, 83% (sex bias = −7.1%) of the forensic sample is correctly classified using the mandible features (method F.3), but 86.3% of the archaeological sample is correctly determined (though the sex bias is higher at −11.9%).

Pooled sample

The same logistic regressions were applied to the pooled sample (Table 4).

Again a 100% accurate model ($\chi^2(41) = 232.683$, $p < 0.001$) (P.1) is obtained when all (excluding the nasal aperture) 14 traits are used. (P.1.1) When traits that by themselves had a high sex bias (≥|20|%) are removed (i.e., mastoid process, external occipital protuberance, nuchal markings, supraorbital margin imprint, zygomatic extension, marginal tubercle, gonial eversion, and gonial angle), the accuracy decreases to 89.3% with a bias of 0.8% ($\chi^2(17) = 133.683$, $p < 0.001$).

When only the cranial traits are used (P.2) ($\chi^2(29) = 153.387$, $p < 0.001$), 90.5% of skulls were correctly classified with a bias of 7.9%.

The model using the mandible (P.3) was also significant ($\chi^2(12) = 109.44$, $p < 0.001$) and correctly classified 85.1% of skulls. The bias was −0.1%.

The temporal traits (P.4) ($\chi^2(6) = 34.958$, $p < 0.001$) had an accuracy of 69.6% and a sex bias of −8.6%. The frontal bone (P.5) resulted in a significant model ($\chi^2(7) = 79.395$, $p < 0.001$) with 80.4% accuracy and −4.5% bias. Walker's five cranial traits (P.6) ($\chi^2(19) = 125.279$, $p < 0.001$) are 86.3% accurate and have a bias of 9.3%. When the mental eminence is removed (P.7) ($\chi^2(15) = 116.118$, $p < 0.001$), the accuracy increases minimally (86.9%) and the bias decreases (5.7%).

When a stepwise logistic regression is used on the whole skull (P.I) ($\chi^2(15) = 149.067$, $p < 0.001$), 90.5% of skulls are correctly classified with a bias of 5.4% when following traits are used: glabella, external occipital protuberance, rugosity, chin shape, mandibular ramus flexure, and gonial angle. The stepwise method on the cranium (P.II) ($\chi^2(9) = 125.193$, $p < 0.001$) utilizes the glabella, external occipital protuberance, and rugosity. This produces an accuracy of 87.5% and a bias of −0.3%. For the mandible (P.III) ($\chi^2(8) = 108$, $p < 0.001$), chin shape, mandibular ramus flexure, gonial eversion, and gonial angle are used. The accuracy using these traits is 84.5% and the bias is −3.6%.

4 | DISCUSSION

4.1 | Intraobserver error test

While initially the intraobserver test showed variable results in terms of perfect agreement, the results improved when additional observation criteria were established. Most traits ended up displaying at least 70% perfect intraobserver agreement, with the only exceptions being the tactile supraorbital margin, mental eminence, and chin shape. Those last two, however, still displayed adequate rates of approximate agreement.

The tactile supraorbital margin is the only trait with questionable intraobserver reproducibility but was used within this research because it might still prove a useful feature in conjunction with other traits.

4.2 | Differences between males and females

Nearly all skull traits included in this study showed a significant difference ($p < 0.05$) in score between males and females, according to the Mann–Whitney *U* test. Only the supraorbital margin (tactile), zygomatic extension, and marginal tubercle did not exhibit this in the archaeological sample. In the forensic and the pooled samples, on the other hand, there was a significant difference. This indicates that the selected characteristics are all somewhat sexually dimorphic, though their usefulness in a sex estimation method is still not proven.

The scoring system of 1 to 5 is often not necessary for the Flemish population, as there are only few individuals with a score of 5. The median of the men is usually a 2 or 3 (see Tables S3–S5). The median of the women for these characteristics is usually 1 or 2. This finding is in line with Walker (2008) and Garvin et al. (2014), who found that the cut point between male and female was usually between 2 or 3 for their mixed samples (which included White, African, and Native Americans, as well as British and Nubian populations). MW considered refining the scores from 1 to 3 into narrower categories, but considering many traits discussed here also use a male/undetermined/female scale, it seemed to overcomplicate the observations rather than refine the results.

It is often said that Dutch female skeletons are very masculine, especially in the lower jaw (Maat et al., 1997). However, the Flemish skulls used in this study indicate that Flemish women are quite gracile and that it is the men who are not hyperrobust. Within the Flemish population, the overlap between males and females is therefore more the result of non-hyper-robust males and not of robust females.

4.3 | Univariate approach

From the univariate analyses, it is clear that a single trait is not ideal for estimating the sex of an individual.

For the archaeological sample, 65% to 83% of the individuals were correctly classified. For the forensic sample, this was 60% to 81%. The pooled sample had an accuracy between 60% and 79% (Table 5). These results are in line with those of Walker (2008), whose univariate analyses yielded accuracies between 69% and 83%, and Garvin et al. (2014), whose results lie between 63% and 79%.

The supraorbital margin imprint had an accuracy between 63.6% and 74.7%, which corresponds to the 67.3% to 74% of the original study (Graw et al., 1999). However, the sex bias (between −47.2% and 45.1%) is too large to be useful (Table 5). Furthermore, the tactile method of the supraorbital margin worked better on the forensic collection than the archaeological one. This is probably due to the better preservation of the orbital margin in the forensic collection. The nasal aperture, considered the best single indicator by Rogers (2005), did not perform very well in this study. It could not be included in the archaeological sample due to inadequate preservation and has a low accuracy and high sex bias in the well-preserved forensic sample (Table 5).

Although some traits achieve high accuracy (≥75%), many also have a very high sex bias (>|10|%). In the archaeological collection, the sex bias is usually positive, meaning more females are correctly classified, while in the forensic collection, the biases are predominantly negative, indicating more males are correctly classified (Table 5). This could be attributed to the fact that the two collections have respectively more women or men in the regression sample and are therefore better at estimating that sex.

TABLE 5 Overview of the percentage of correctly classified individuals and the sex bias based on several individual skull features, for the archeological, forensic and pooled samples. Methods with a correct classification rate ≥75% and a sex bias ≤|10|% are in bold. Sex bias = % correctly classified females − % correctly classified males.

	Archeological		Forensic		Pooled	
Trait	Correctly classified (%)	Sex bias (%)	Correctly classified (%)	Sex bias (%)	Correctly classified (%)	Sex bias (%)
Glabella	**76.1**	**−4.7**	**80.2**	**−4.2**	**78.4**	**−5.2**
Mastoid process	68.4	45.7	63.6	−10.1	65.3	38.9
External occipital protuberance	80	53.8	70.1	−31.1	74.5	27.8
Nuchal markings	**76**	**−4.1**	70.1	−45.1	72.8	−24.5
Supraorbital margin (tactile)	66.7	55.1	72.7	−2.9	64.4	−9.2
Supraorbital margin (imprint)	74.7	45.1	63.6	−47.2	66.9	32.4
Nasal aperture	77.3	50.4	60.5	32.6	64.1	37.4
Zygomatic extension	67.6	96	61.4	30.9	61.1	22
Marginal tubercle	64.8	63	60.2	−31.9	60.4	46.1
Rugosity	77.1	−10.7	80.7	−16.9	79.1	−14.4
Mental eminence	74.3	14.5	70.1	−46.8	69.6	17.2
Chin shape	77	35.9	74.7	−12.4	74.5	−12.7
Mandibular ramus flexure	**76.6**	**5.4**	67	11.7	71.5	15.1
Gonial eversion	70.5	11.7	70.5	−31.4	66.3	−41
Gonial angle	83.1	16.1	65.9	−23.4	73.3	30.7

Only the glabella seems to consistently show a relatively high accuracy (between 76.1% and 80.2%) and a small bias (between −4.7% and −5.2%) compared with the other traits. This is therefore the only trait that, if given no other options, can be used alone for a sex estimation. A score of 2 or higher would indicate a male.

4.4 | Multivariate approach

As expected, the multivariate analyses outperformed the univariate ones. Table 4 provides an overview of the correct classification rates and sex biases of the logistic regression methods discussed in Section 3.3. Furthermore, the same table shows how well the different methods work on the different samples.

For each skull region (full skull, cranium, mandible, temporal, and frontal), the best method was chosen. A method is considered the best if the correct classification is ≥80%, and the sex bias is ≤|10|%, for both the archaeological and forensic samples. In addition, the number of necessary traits is limited to 6 at most.

When all skull traits are used in a logistic regression, a correct classification of 100% can, theoretically, be achieved (Table 4; A.1, F.1, and P.1). If only the cranial features or the mandible are used (regressions A.2, F.2, F.2.1, P.2, A.3, F.3, and P.3), there is still a very high correct classification, with a small bias. However, in practice, this will rarely be feasible due to insufficient preservation.

For the archaeological sample, a 100% correct classification could be obtained with only five of the 14 variables, those being the glabella, nuchal markings, mental eminence, gonial eversion, and gonial angle (A.1.1). However, this method did not work as well on the forensic sample, with only 76.1% of skulls being correctly classified.

In contrast to the archaeological formulas rarely working well on the forensic skulls, the forensic formulas seem to partially work on the archaeological skulls. This may be due to the higher reliability of the forensic sample, as those sexes are known and not estimated based on the pelvis like in the archaeological sample. Furthermore, differences in sexual dimorphism over time may also play a role, due to differences in lifestyle, diet, and other environmental factors. The cranium seems to show the most differences between the two collections, while the frontal bone shows the least. This could indicate that the forehead has undergone the fewest temporal changes and might be less influenced by environmental factors than other parts of the skull like the mandible, which is highly influenced by diet (Katz et al., 2017).

The methods that do well in both the archaeological and forensic collections are A.5 and FII, and several methods based on the pooled sample: P.1, P.1.1, P.3, P.5, P.I, P.II, and P.III.

4.4.1 | Skull

For the skull methods, P.1, P.1.1, and P.I work well.

Method P.1 uses 14 skull features, that is, all features except the nasal aperture and has a correct classification of 100%. Though this method works very well, it can only be used on exceptionally well preserved skulls.

In contrast, methods P.1.1 and P.I use only six traits each and can therefore be used more. On the pooled sample, the correct classification rate of P.1.1 is 89.3% and the sex bias is 0.8%. For the archaeological sample alone, this is respectively 92.5% and 3.6% and for the forensic sample 86.4% and −3.8%. P.I has a correct classification rate of 90.5% and a bias of 5.4% for the pooled sample. This is 92.5% and 8.8% for the archaeological sample and 88.6% and 1.5% for the forensic sample.

Method P.1.1[1,2] is recommended because of the smaller sex bias.

With an accuracy of 89.3% this method achieves similar results to one of Walker's (2008) methods, which had an accuracy of 88% for a pooled sample of an American (White and Black) and English population, and Garvin et al. (2014), who achieved 86.1% accuracy for their pooled US Whites sample.

4.4.2 | Cranium

The two methods that consistently work well on the cranium are F.II and P.II. Both are stepwise logistic regressions and use the glabella, external occipital protuberance, and rugosity. The percentage of correctly classified individuals and the sex bias are also the same for both methods: 87.5% and −0.3% respectively for the pooled sample, 85% and 1.8% for the archaeological sample, and 89.8% and −0.5% for the forensic sample. So both formulas can be used.

4.4.3 | Mandible

Though the mandible features always perform poorer than the cranial features, the difference is not as big as in the study by Maat et al. (1997). In that study on Dutch skeletons, the sex of the cranium matched that of the pelvis in 95.7% of the cases; for the mandible, this was 69.5%, with more than half of the females being incorrectly classified. In this study, 78% to 89% of the mandibles are correctly classified with a much smaller sex bias between −8.3% and 11.2% (Table 4, A.3, A.III, F.3, F. III, P.3, and P.III). These different results may indicate a population difference, with Dutch women having more robust mandibles than Flemish women. But this could also be the result of differences in methodology, as Maat et al. (1997) used only the mental eminence, the gonial angle, and the robustness of the mandible and inferior margin.

Methods P.3 and P.III work well for both samples, though P.3 is recommended because of its higher accuracy and lower bias.

[1]First, the predicted log odds must be calculated. For this, enter either 1 or 0 in the place of the relevant skull feature. Fill in 1 with the correct score and 0 with the other scores of that feature. For example, for a glabella with a score of 2 this would be: 3.573 + (−0.789 * 0) + (0.788 * 1) + (2.219 * 0) + …. Then, the predicted probability can be calculated. *e* (Euler's number) is a mathematical constant equal to 2.71828, which can be found on most (scientific) calculators. In SPSS following calculation can also be made instead: "= 1/(1 + EXP (−[predicted log odds]))."

[2]Here, the equations include square brackets [] for readability. In SPSS, these have to be replaced by round brackets ().

P.1.1 (glabella, supraorbital margin tactile, rugosity, mental eminence, chin shape, & mandibular ramus flexure)

Accuracy (pooled sample): 89.3%

Sex bias (pooled sample): 0.8%

Predicted log odds = 3.573 + (-0.789 * [glabella=1]) + (0.788 * [glabella=2]) + (2.219 * [glabella=3]) + (0.984 * [supraorbital margin tactile=1]) + (0.094 * [supraorbital margin tactile=2]) + (0.513 * [supraorbital margin tactile=3]) + (0.697 * [supraorbital margin tactile=4]) + (-3.511 * [rugosity=1]) + (-1.579 * [rugosity=2]) + (-1.229 * [mental eminence=1]) + (-0.974 * [mental eminence=2]) + (-0.606 * [mental eminence=3]) + (0.29 * [mental eminence=4]) + (-1.722 * [chin shape=1]) + (-0.28 * [chin shape=2]) + (-0.5 * [mandibular ramus flexure=1]) + (-3.005 * [mandibular ramus flexure=2])

Predicted probability = $1/(1+e^{(-[\text{predicted log odds}])})$

Predicted probability ≥0.5 = male

Predicted probability <0.5 = female

F.II (glabella, external occipital protuberance, & rugosity)

Accuracy (pooled sample): 87.5%

Sex bias (pooled sample): -0.3%

Predicted log odds = 20.652 + (-0.479 * [glabella=1]) + (0.621 * [glabella=2]) + (20.799 * [glabella=3]) + (-19.254 * [external occipital protuberance=1]) + (-18.757 * [external occipital protuberance=2]) + (-15.739 * [external occipital protuberance=3]) + (-19.526 * [external occipital protuberance=4]) + (-3.492 * [rugosity=1]) + (-1.598 * [rugosity=2])

Predicted probability = $1/(1+e^{(-[\text{predicted log odds}])})$

Predicted probability ≥0.5 = male

Predicted probability <0.5 = female

P.3 (mental eminence, chin shape, mandibular ramus flexure, gonial eversion, & gonial angle)

Accuracy (pooled sample): 85.1%

Sex bias (pooled sample): -0.1%

Predicted log odds = 3.976 + (-0.736 * [mental eminence=1]) + (-0.29 * [mental eminence=2]) + (0.116 * [mental eminence=3]) + (0.016 * [mental eminence=4]) + (-2.644 * [chin shape=1]) + (-0.987 * [chin shape=2]) + (-0.712 * [mandibular ramus flexure=1]) + (-2.91 * [mandibular ramus flexure=2]) + (-1.844 * [gonial eversion=1]) + (-0.113* [gonial eversion=2]) + (-1.869 * [gonial angle=1]) + (-1.786 * [gonial angle=2])

Predicted probability = 1/(1+e$^{(\text{-[predicted log odds]})}$)

Predicted probability ≥0.5 = male

Predicted probability <0.5 = female

F.4 (mastoid process & zygomatic extension)

Accuracy (pooled sample): 69.6%

Sex bias (pooled sample): -8.6%

Predicted log odds = 22.045 + (-41.761 * [mastoid process=1]) + (-20.837 * [mastoid process=2]) + (-21.243 * [mastoid process=3]) + (-19.949 * [mastoid process=4]) + (-1.487 * [zygomatic extension=1]) + (0.372 * [zygomatic extension=2])

Predicted probability = 1/(1+e$^{(\text{-[predicted log odds]})}$)

Predicted probability ≥0.5 = male

Predicted probability <0.5 = female

A.5 (glabella & supraorbital margin tactile)

Accuracy (pooled sample): 80.4%

Sex bias (pooled sample): -4.6%

Predicted log odds = 1.608 + (-2.754 * [glabella=1]) + (-0.084 * [glabella=2]) + (0.916 * [glabella=3]) + (-0.889 * [supraorbital margin tactile=1]) + (-1.049 * [supraorbital margin tactile=2]) + (-0.919 * [supraorbital margin tactile=3]) + (-0.425 * [supraorbital margin tactile=4])

Predicted probability = $1/(1+e^{(-[\text{predicted log odds}])})$

Predicted probability ≥0.5 = male

Predicted probability <0.5 = female

4.4.4 | Temporal

The method for the temporal bone does not work well within the archaeological sample (A.4). The correct classification rate was quite low while the sex bias was very high, 70% and 30%, respectively. In the forensic collection (F.4), the sex bias improved (−8.2%), but the correct classification remained low (70%). Therefore, this bone is not recommended for use in sex estimation.

However, when dealing with a collection of highly fragmented skulls, one may be forced to apply this method. In such cases, the use of method F.4 or P.4 is recommended as the accuracy approaches 70% and the bias is ≤|10|%.

4.4.5 | Frontal

For the frontal bone, A.5 and P.5 work well with an accuracy of 80.4% and a bias of −4.6% for the pooled sample. For the archaeological sample, this is 80% and −6.3%; while for the forensic sample, this is 80.7% and −3.1%.

4.5 | Limitations of the study

The main limitation of this study is that the genetic sex of the archaeological sample is unknown and had to be estimated by observations of the pelvic bones. Though this is considered the best morphological sexing method, this undoubtedly introduced some bias. However, molecular sex determination methods were out of the scope of this study. Testing the method on another known sex sample may be useful in the future but was beyond the scope of this project.

Another factor that potentially adds bias are the missing values in the dataset due to unobservable traits (usually due to fragmented and damaged skeletons in the archaeological sample). Unfortunately, missing values are unavoidable when working with archaeological samples, and when the total sample was considered, in this study, median imputation was used to achieve a complete dataset. We also addressed the issue of missing values by using a smaller number of traits for some of the analyses (e.g., only frontal bone) to make the method widely applicable to fragmented human remains.

5 | CONCLUSION

Various logistic regressions were performed on two Flemish skeletal collections using up to 15 skull features to investigate how reliable the morphology of the skull in the Flemish population is to estimate sex in an archaeological context. The test sample consisted of 168 skulls from two collections: an archaeological one (Aalst, mid-15th to 17th century) and a known-sex forensic collection (East-Flanders, 1930s–1990s).

The univariate logistic regressions resulted in correct classification rates between 60% and 83%. However, the sex bias is often too high (>|10|%) to be useful. The only solo trait that could potentially be useful is the glabella (accuracy: 78.4% and sex bias −5.2% for the pooled sample). According to this trait, any Flemish individual with a glabella score ≥2 should be classified as male.

The use of more features benefits the overall correct classification rates, as shown by the various multivariate logistic regressions. When all traits are used, a 100% correct classification can be obtained. However, using 14 or 15 features is not always possible in practice. Hence, there is a need for additional or alternative methods that use fewer or different combinations of features. The different sex estimation equations derived from the logistic regression models were therefore tested on the two collections, and the best methods were selected. Accuracy (minimum 80%), sex bias (maximum |10|%), and usefulness on poorly preserved and/or incomplete skulls were taken into account.

For a whole skull, method P.1.1 is recommended. This method uses the glabella, supraorbital margin (tactile), rugosity, mental eminence, chin shape, and mandibular ramus flexure. The accuracy for the pooled sample is 89.3% with a very low sex bias of 0.8%. For a cranium, that is, a skull without a mandible, methods F.II and P.II are recommended. These use the glabella, external occipital protuberance, and rugosity of the cranium. The accuracy is 87.5% and the bias −0.3% for the pooled sample. When only a mandible is available, method P.3 can be used, using the mental eminence, chin shape, mandibular ramus flexure, gonial eversion, and gonial angle; 85.1% of individuals (pooled sample) were correctly classified with a very low bias of −0.1%. For the temporal bone, method F.4 or P.4 (mastoid process and zygomatic extension) can be used, though these methods have a lower accuracy (69.6%) and greater bias (−8.6%) than the others. A frontal bone (A.5 and P.5; glabella and supraorbital margin tactile) can be used to estimate sex with 80.4% accuracy and −4.6% bias.

Because these results are promising, it would be constructive to explore the subject further. The need for the popular 1 to 5 scoring system should be questioned, as the male median in this study never exceeded a score of 3. Furthermore, the methods should be tested on other European populations. Flemish mandibles, for example, seem to differ from Dutch ones despite the close geographical proximity of the two populations. This further highlights the need for population specific formulae. More research should be done to help investigate these population differences as currently it can be quite difficult to compare studies due to different approaches and reporting (e.g., sex bias is not always reported).

To summarize, we can conclude that the skull is sexually dimorphic enough in various aspects to be used as a sex indicator in a Flemish population. For the whole skull, the cranium, the mandible, and the frontal, sex discriminant formulas can be developed that have an accuracy of at least 80% and a sex bias of at most |10|%. Furthermore, a small, specific part of the skull, namely, the eyebrows (i.e., frontal bone with an intact glabella and an intact orbit), can be used on its own for sex estimation.

ACKNOWLEDGMENTS

We would like to thank the Ghent University Museum for granting us access to their known-sex collection of human skulls. We are also grateful to the reviewers for their useful comments and S. J. Patterson for proofreading.

CONFLICT OF INTEREST STATEMENT

The authors have no conflicts of interest to declare.

DATA AVAILABILITY STATEMENT

The raw data are available upon reasonable request from the lead author.

ORCID

Maggie Wongsantativanich https://orcid.org/0009-0004-6032-7694

Isabelle De Groote https://orcid.org/0000-0002-9860-0180

REFERENCES

Boucherie, A., Polet, C., Lefèvre, P., & Vercauteren, M. (2021). Sexing the bony labyrinth: A morphometric investigation in a subadult and adult Belgian identified sample. *Journal of Forensic Science*, *66*(3), 808–820. https://doi.org/10.1111/1556-4029.14663

Boudin, M., Bru, M.-A., De Clercq, G., Dierkens, A., Laleman, M. C., Quintelier, K., Vandenbruaene, Van Strydonck, M., & Vermeiren, G. (2009). *Onder het Sint-Pietersplein Gent. Van hoogadellijke begraafplaats tot parking*. Uitgeverij Snoeck.

Bruggeman, J., Claessens, L., Coremans, L., Hellinx, A.-J., Reyns, N., & Van Buggenhout, J. (2021). 10 jaar. 10 verhalen – 10 foto's. Rapporten All-Archeo bvba, 1000. Retrieved from All-Archeo website https://www.all-archeo.be/rapporten/1000_10%20jaar_verhalen/Rapport1000VerhalenFotos.pdf

Bruggeman, J., Coremans, L., Van Buggenhout, J., Claessens, L., Ferket, R., & Reyns, N. (2019). Archeologische opgraving Aalst – Louis D'Haeseleerstraat. *Annuntiaten- en* Theresianenklooster. Rapporten All-Archeo bvba, 316. Retrieved from Onroerend Erfgoed website https://oar.onroerenderfgoed.be/publicaties/ROEV/4041/ROEV4041-001.pdf

Bruzek, J., Sefcakova, A., & Cerny, V. (2004). Révision du sexe des squelettes épipaléolithiques de Taforalt et d'Afalou-bou-Rhoummel par une approche probabiliste. *Antropo*, *7*, 195–202.

Buikstra, J. E., & Mielke, J. H. (1985). Demography, diet and health. In R. I. Gilbert & J. H. Mielke (Eds.), *The analysis of prehistoric diets* (pp. 359–422). Academic Press.

Buikstra, J. E., & Ubelaker, D. H. (1994). Standards for data collections from human skeletal remains. Arkansas Archaeological Survey.

Burns, K. R. (2012). *Forensic anthropology training manual* (3rd ed.). Routledge.

Capitaneanu, C., Willems, G., Jacobs, R., Fieuws, S., & Thevissen, P. (2017). Sex estimation based on tooth measurements using panoramic radiographs. *International Journal of Legal Medicine*, *131*, 813–821. https://doi.org/10.1007/s00414-016-1434-0

Christensen, A. M., Passalacqua, N. V., & Bartelink, E. J. (2014). *Forensic anthropology. Current methods and practice*. Academic Press.

De Groote, I., & Humphrey, L. T. (2016). Characterizing evulsion in the later stone age Maghreb: Age, sex and effects on mastication. *Quaternary International*, *413*, 50–61. https://doi.org/10.1016/j.quaint.2015.08.082

De Groote, I., Van de Vijver, K., Veselka, B., De Potter, P., Massagé, L., Van der Dooren, L., Vandenborre, J., Larmuseau, M. H. D., Danckers, J., & Robberechts, B. (2023). MEMOR: A database of archeological human remains collections from Flanders, Belgium. *American Journal of Biological Anthropology*, *181*, 677–681. https://doi.org/10.1002/ajpa.24801

Garvin, H. M., Sholts, S. B., & Mosca, L. A. (2014). Sexual dimorphism in human cranial trait scores: Effects of population, age, and body size. *American Journal of Physical Anthropology*, *154*(2), 259–269. https://doi.org/10.1002/ajpa.22502

Graw, M., Czarnetzki, A., & Haffner, H. T. (1999). The form of the supraorbital margin as a criterion in identification of sex from the skull: Investigations based on modern human skulls. *American Journal of Physical Anthropology*, *108*(1), 91–96. https://doi.org/10.1002/(SICI)1096-8644(199901)108:1<91::AID-AJPA5>3.0.CO;2-X

Kajanoja, P. (1966). Sex determination of Finnish crania by discriminant function analysis. *American Journal of Physical Anthropology*, *24*(1), 29–33. https://doi.org/10.1002/ajpa.1330240104

Kamnikar, K. R., Hefner, J. T., Monsalve, T., & Florez, L. M. B. (2022). Craniometric variation in a regional sample from Antioquia, Medellin, Colombia: Implications for forensic work in the Americas. *Forensic Anthropology*, *5*(3), 239–250. https://doi.org/10.5744/fa.2020.2023a

Katz, D. C., Grote, M. N., & Weaver, T. D. (2017). Changes in human skull morphology across the agricultural transition are consistent with softer diets in preindustrial farming groups. *Proceedings of the National Academy of Sciences of the United States of America*, *14*(34), 9050–9055. https://doi.org/10.1073/pnas.1702586114

Kiskira, C., Eliopoulos, C., Vanna, V., & Manolis, S. K. (2022). Biometric sex assessment from the femur and tibia in a modern Greek population. *Legal Medicine*, *59*, 102126. https://doi.org/10.1016/j.legalmed.2022.102126

Klales, A. R. (2020). Sex estimation using pelvis morhpology. In A. R. Klales (Ed.), *Sex estimation of the human skeleton* (pp. 75–93). Academic Press. https://doi.org/10.1016/B978-0-12-815767-1.00006-7

Loth, S. R., & Henneberg, M. (1996). Mandibular ramus flexure: A new morphologic indicator of sexual dimorphism in the human skeleton. *American Journal of Physical Anthropology*, *99*(3), 473–485. https://doi.org/10.1002/(SICI)1096-8644(199603)99:3<473::AID-AJPA8>3.0.CO;2-X

Maat, G. J., Mastwijk, R. W., & van der Velde, E. A. (1997). On the reliability of non-metrical morphological sex determination of the skull compared with that of the pelvis in The Low Countries. *International Journal of Osteoarchaeology*, *7*(6), 575–580. https://doi.org/10.1002/(SICI)1099-1212(199711/12)7:6<575::AID-OA308>3.0.CO;2-4

Mays, S., & Cox, M. (2000). Sex determination in skeletal remains. In M. Cox & S. Mays (Eds.), *Human osteology in archaeology and forensic science* (pp. 117–130). Cambridge University Press.

Phenice, T. W. (1969). A newly developed visual method of sexing the Os pubis. *American Journal of Physical Anthropology*, *30*(2), 297–302. https://doi.org/10.1002/ajpa.1330300214

Pickering, R., & Bachman, D. (2009). *The use of forensic anthropology* (2nd ed.). CRC Press.

Pons, J. (1955). The sexual diagnosis of isolated bones of the skeleton. *Human Biology*, *27*(1), 12–21. http://www.jstor.org/stable/41448115

Relethford, J. H. (2004). Global patterns of isolation by distance based on genetic and morphological data. *Human Biology*, *76*(4), 499–513. https://doi.org/10.1353/hub.2004.0060

Rogers, T. L. (2005). Determining the sex of human remains through cranial morphology. *Journal of Forensic Science*, *50*(3), 493–500. https://doi.org/10.1520/JFS2003385

Silva, J., Araya, C., Pardo, S., Reyes, T., Salcedo, A., Sanhueza, A., & Liberona, S. (2020). Tubérculo Marginal del Hueso Cigomátic. *International Journal of Morphology*, *38*(1), 159–164. https://doi.org/10.4067/S0717-95022020000100159

Šlaus, M., & Tomičić, Ž. (2005). Discriminant function sexing of fragmentary and complete tibiae from medieval Croatian sites. *Forensic Science International*, *147*(2–3), 147–152. https://doi.org/10.1016/j.forsciint.2004.09.073

Van den Eynden, J., Descamps, T., Delporte, E., Roosens, N. H. C., De Keersmaecker, S. C. J., De Wit, V., Vermeesch, J. R., Goetghebeur, E., Tafforeau, J., Demarest, S., Van den Bulcke, M., & Van Oyen, H. (2018). The genetic structure of the Belgian population. *Human Genomics*, *12*(6), 1–9.

Walker, P. L. (2008). Sexing skulls using discriminant function analysis of visually assessed traits. *American Journal of Physical Anthropology*, *136*(1), 39–50. https://doi.org/10.1002/ajpa.20776

Walrath, D. E., Turner, P., & Bruzek, J. (2004). Reliability test of the visual assessment of cranial traits for sex determination. *American Journal of Physical Anthropology*, *125*(2), 132–137. https://doi.org/10.1002/ajpa.10373

White, T. D., Black, M. T., & Folkens, P. A. (2012). *Human osteology* (3rd ed.). Academic Press.

White, T. D., & Folkens, P. A. (2005). *The human bone manual*. Elsevier Academic Press.

Williams, B. A., & Rogers, T. L. (2006). Evaluating the accuracy and precision of cranial morphological traits for sex determination. *Journal of Forensic Sciences*, *51*(4), 729–735. https://doi.org/10.1111/j.1556-4029.2006.00177.x

SUPPORTING INFORMATION

Additional supporting information can be found online in the Supporting Information section at the end of this article.

How to cite this article: Wongsantativanich, M., & De Groote, I. (2024). Nonmetric population-specific sex estimation based on the skull using logistic regression for Flemish samples. *International Journal of Osteoarchaeology*, *34*(4), e3308. https://doi.org/10.1002/oa.3308