***An approach to establishing a workflow pipeline for synergistic analysis of osteological and biochemical data. The case study of Amvrakia in the context of Corinthian colonisation between 625-189 BC in Epirus, Greece.***

Xanthopoulos Kiriakos\*1, Georgiadou Aggeliki1, Papageorgopoulou Christina1

1Laboratory of Physical Anthropology (Democritus University of Thrace) – Komotini, Greece

\*Corresponding author

Correspondence: ancbiosoc@gmail.com

**Abstract**

Bioarchaeology has long focused on understanding past human life through skeletal remains, including oral pathology and stable isotope analysis. Despite advancements in statistical analysis, correlations are still largely made manually. To streamline this, we propose a tool to automatically identify correlations between dental pathologies and isotope values. Our work is part of the ongoing APOIKIA project on Corinthian colonisation in 7th century BC. We analyze osteoarchaeological material from the Western Necropolis of Amvrakia, Greece, founded around 640 BC. Our methodology integrates novel data collection approaches and FAIR principles for data longevity and accessibility, utilizing R scripts and cloud webware.

***Keywords:*** Osteoarchaeology, Colonisation, Data analysis, Archaeological methodology

**Introduction**

Bioarchaeology has laboured for many decades to reconstruct the living conditions and socioeconomic aspects of past human experiences through the study of skeletal remains. In lack of either material or textual evidence, bioarchaeological approaches hold great value in developing a synergy effect with historical, demographic and archaeological data. Within the vastly broad area of bioarchaeological research lies oral pathology, and stable carbon and nitrogen analysis. These approaches are regarded today as powerful tools for archaeologists to monitor the shifts in the carbohydrate intake and the dietary protein consumption of past populations. Several previous studies have demonstrated the combination of stable isotopic and dental health data to shed light on the lifestyle, diet, nutrition, and subsistence economies(White, 2007; Ossenberg, 2013; Colleter et al., 2020; Constantinou & Nikita, 2022; Caruso et al., 2023; Georgiadou et al., 2023; Caruso & Nikita, 2024). Despite the significant efficiency in the correlation process of bioarchaeological data through varied statistical analyses, the estimations are still performed by the observation, interpretation and comparison of different statistical plots. To avoid time-consuming procedures, we are in the process of generating a tool that enables the researcher to automatically find if there are any correlations between dental pathologies and isotope values.

Our research is part of an ongoing project entitled “APOIKIA: Ancient DNA analysis in novel multidisciplinary approach of ancient Corinthian colonisation. Ancient Amvrakia and Ancient Tenea as demonstration examples”. Amvrakia (fig. 1) was officially founded in 625 BC during various colonisation endeavours of the Corinthians in Epirus, Greece. Five research institutes have partnered to enhance the archaeological research on the phenomenon of colonisation by elaborating current lines on the scientific network or carving new ones. The study presented here focuses on the osteoarchaeological and biochemical research field of the project.

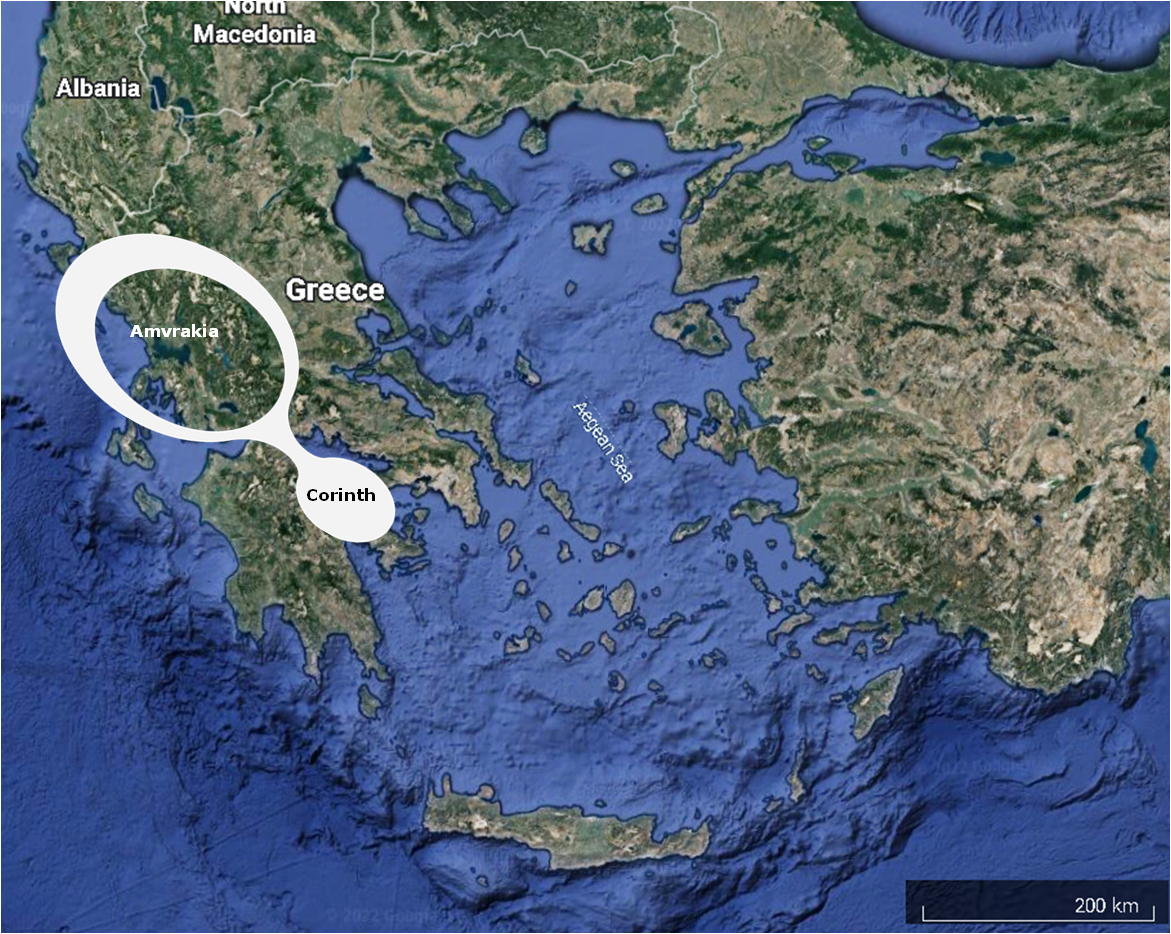
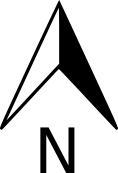


Figure 1 Corinth(metropolis) and Amvrakia(colony) in Greece

Bioarchaeologists are all the more specialising in a specific field of study as technological advances limit the researchers’ ability to keep track of them. Collaboration among specialists is the only viable way to conduct research in the coming years. The obstacles that arise through such collaborations may throw whole scientific projects off balance or even lead to their termination. One of those obstacles is the complexity of merging diverse types of recorded data into a unified database. In attempts to address this challenge, bioarchaeologists have turned to utilizing different web-based platforms, including Google Drive, Dropbox, Box, and GitHub. While these tools have been beneficial, they often lack a cohesive interface, making them less user-friendly. To mitigate these issues, there is a growing trend within the scientific community to develop new tools, ranging from mobile applications to cloud-based software, with varying licenses and methodological applications, spanning historical research to forensic law enforcement. (Bartholdy, Hoogland, and Waters‐Rist 2019; Berg and Kenyhercz 2017; Colleter, Romain, and Barreau 2020; Hunt 2019; Klales 2020; G. R. Scott et al. 2018).

**Material and Methods**

The osteoarchaeological material comes from various excavational periods of the Western Necropolis (WN) and Eastern Necropolis (EN) of Amvrakia that took place between 1993-2022 (fig. 2).   
The WN stands out as one of Greece's largest and most intact necropolises uncovered to date. It has revealed over 600 grave sites, collectively containing more than 700 individuals. The EN of Amvrakia is more segmented than WN with approximately 200 grave sites and around 100 individuals. The fragmented state of skeletal preservation in both Necropolises posed an interesting challenge of creating a proper database for commingled and individual burial (Adams & Konigsberg, 2008) as well as establishing a convenient methodology for filtering and selecting the most appropriate samples for isotope analysis.



Figure 2 Western Necropolis (white box 1), Eastern Necropolis (white box 2), City of Amvrakia (inside the red perimeter) (Papadopoulou et al., 2021)

The methodology was designed on the basis of novel approaches to data collection and FAIR (Findable, Accessible, Interoperable, Reusable) principles (Schmidt and Marwick 2020). Proprietary software is used as minimally as possible and only for transitioning from old but established methodologies into new tools based on free and preferably open-source software. The principle behind this is to maximise the longevity of data and maintain their expediency. In particular, we create R programming language scripts that correlate paleodemographic, palaeopathological, osteometric and isotope data by following revised methodology and utilising free cloud webware. To adhere to each tenet of FAIR principles, meticulous scrutiny was applied to the processes of material management and data documentation.

First of all, findability of osteological material was facilitated through the implementation of a meticulously devised file storage protocol, informed by internationally recognized guidelines (Buikstra & Ubelaker, 1994; Schwartz, 1995; Scheuer & Black, 2000; Hillson, 2005; Lewis, 2007; Schaefer et al., 2009; White et al., 2012; Nikita, 2017; Seguchi & Dudzik, 2019; Nikita & Karligkioti, 2020). All files were systematically recorded and stored locally, leveraging the cloud-based service provided by Dropbox. This approach not only ensures accessibility but also offers robust mechanisms for file integrity preservation, enabling restoration to previous versions in cases of file corruption. Additionally, the platform extends the option to share files publicly, enhancing accessibility and dissemination.

Data pertaining to storage were meticulously documented in spreadsheets, utilizing file formats such as .ods or .xlsx. Demographic, pathological, and osteometric data, on the other hand, were systematically recorded in databases, employing file formats like .odb or .accdb. To facilitate statistical analyses, databases were exported to spreadsheets in the .csv format, ensuring compatibility and ease of manipulation.

The skeletal material was securely stored within the infrastructure of the archaeological museum in Arta (Greece) to facilitate accessibility for researchers. To further ensure accessibility, we have proposed the use of governmental servers so that future researchers can have remote access.   
Despite the fact that the Greek legal framework has not yet addressed the matter, there are indications that it will do so in the near future(Hellenic Republic, 2015).

The selection of file types and platforms, whether local or cloud-based, facilitates interoperability. Given the swift pace of technological advancement, it is essential to employ files and platforms that ensure medium-term data usage across diverse computational systems. This approach remains crucial until novel technologies offer alternative solutions to preserve interoperability.

The utilization of foundational file formats –such as .csv-, particularly in conjunction with script programming in R (Marwick, 2017), facilitates the reusability of both data analysis procedures and the raw data itself. The R programming language safeguards the sustainability of data analysis processes, even amidst alterations in the file formats currently employed for statistical analysis brought about by technological advancements (RStudio Team, 2020). Leveraging R programming packages capable of seamlessly converting file formats enhances the adaptability of data files for future applications.

+++

We create databases based on international bibliography and protocols created in the Laboratory of Physical Anthropology of Democritus University of Thrace. We use macroscopy to record the morphoscopical indices -for age calculation and sex determination- and pathological lesions –for palaeopathology study). We use endoscopy to record specific pathologic lesions –eg Hyperostosis Frontali Interna- in the endocranium. We use digital osteometric tools to record dimensions of bone that can faciliatate height calculation.

We create a pipeline of procedures that begin with the recording of bioarchaeological and biochemical data and finish with their statistical analysis while valuable insights are generated alongside a webware environment. We use open-source databases and spreadsheets like Open Office Database and Calc and Dropbox as software and digital repositories items for feeding data and R (Marwick 2017, 432) as the programming language for creating webware applications that process the data (RStudio Team 2020). We utilize R packages[[1]](#footnote-2) to facilitate database management, data tidying, and statistical analyses in addition to the default packages provided by the R programming language.

Statistical analysis follows the standard methodology employed in bioarchaeological studies. Descriptive statistical analysis aims to gain a general overview of the data and explore potential patterns in its distribution. Specifically, we analyze the distribution of age groups and biological sex. Additionally, bone measurements are examined to compute measures such as mean, median, and dispersion (overall and by biological sex). Furthermore, the frequency and distributions of dental lesions (overall, by biological sex, and age), osteoarthritic lesions (overall, by age and by biological sex), as well as cranial and post-cranial lesions (overall, by biological sex and by age) are analyzed.

Based on the results of descriptive statistics, more elaborate statistical analyses is employed. Exploratory Data Analysis is conducted in order to investigate the correlation between bone measurements and age by using techniques such as scatterplots, histograms, and correlation matrices (Vark & Howells, 1984).

Predictive Modeling is anticipated to be employed only if the samples and data are considered sufficient and reliable. In this case, the correlation of bone measurements for biological sex inference is examined through regression analysis (Viciano et al., 2013). Hypothesis testing, including t-tests, chi-square tests, ANOVA, and correlation tests, is also conducted to test correlations between parameters and unravel differences between groups e.g. between males and females (Nikita et al., 2013). Bayesian statistics is employed to explore age of death based on new data derived from prior knowledge. This field provides an insight into the data, enhancing parameter interpretation (Hens & Godde, 2020). Survival Analysis is utilized to estimate the risk of disease occurrence between groups. The time to event occurrence, such as the time between disease onset and individual death, is estimated using techniques such as Kaplan-Meier estimation and Cox proportional hazards regression (Kubehl & Temple, 2020). Multivariate Analysis is applied to examine correlations between morphological and metric characteristics and to identify patterns (Calce et al., 2017). Techniques to be applied include Principal Component Analysis, Factor Analysis, and Cluster Analysis.

**Recording Data**

Prior to commencing data recording, researchers must deliberate on the approach to be undertaken. This decision, as depicted in Figure 3, hinges primarily on the physical proximity of collaborators, which is contingent upon the accessibility of the study material.

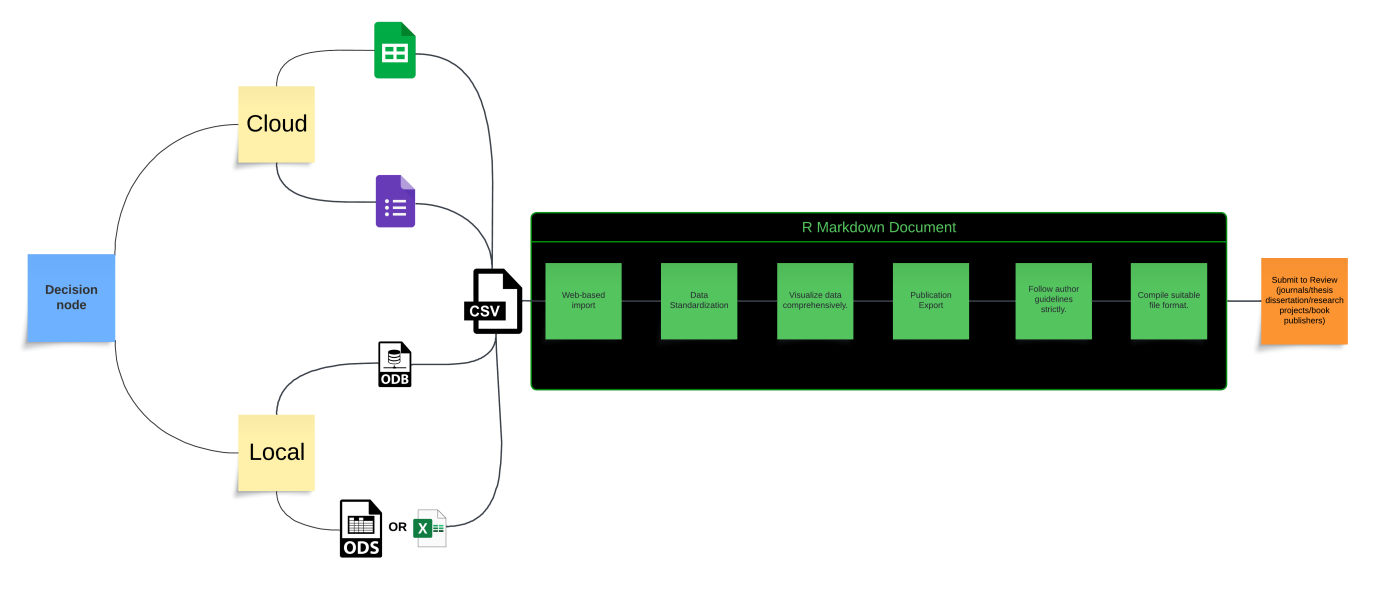


Figure 3 Design of workflow pipeline

In one scenario, Google Cloud environment is employed due to its provision of diverse tools facilitating simultaneous collaboration. Of particular significance is the integrated chat room feature, enabling real-time communication and swift resolution of data wrangling issues, thereby optimizing time efficiency. The offline functionality is invaluable in situations where researchers opt to work within a browser environment, mitigating the risk of data loss in the event of internet connectivity disruptions. Upon reconnecting to the internet, cloud files are automatically updated, ensuring seamless collaboration and synchronization of data across collaborators' files.

Alternatively, researchers may opt for Open Office spreadsheets and/or locally saved databases on Dropbox. In this setup, researchers work with their respective local databases and subsequently import them into the web-based application.

For researchers desiring bespoke cloud services, the option exists to leverage open-source technologies underlying cloud server environments such as ~okeanos (https://okeanos.grnet.gr/). These “Infrastructures as a Service” (IaaS) can facilitate even further the online data collection in cases of remote research teams and collaborators.

After the decision is made, the researchers start gathering the data including demographic information, pathological indices, osteometric measurements, stable isotope and ancDNA samples.

**Importing and Tidying Datasets**

Datasets are uploaded onto the web-based application via drag-and-drop or local upload mechanisms. The application facilitates data table amalgamation while allowing for specific standardization and data cleaning techniques such as NA omission and laterality determination for bilateral indices. Parameterization protocols adhere to standard statistical analysis methodologies derived from pertinent literature pertaining to the research domain. In our case study, parameterization is informed by literature on age estimation, sex determination, height estimation, dental pathology, and isotope analysis[[2]](#footnote-3).

After thorough data cleaning is conducted, the data are preprocessed to identify and rectify any errors, inconsistencies, or missing values in the datasets. Quality control measures are implemented to assess the reliability and validity of the collected data, including checks for sample contamination, data outliers, and measurement errors (Sokal & Rohlf, 2009; Baxter, 2012; Nikita, 2020).

**Analyses and Publication**

Following parameter definition, the application generates descriptive tables and correlation plots. Subsequently, researchers engage in result discussions and formulate conclusions that inform the formulation of new research inquiries. Relevant features and variables from both osteological and biochemical datasets are identified during this process in order to address the research questions and/or hypotheses. Statistical analyses explore the relationships among osteological and biochemical variables and the findings are further interpreted by researchers so that implicit biological or social mechanisms can be discovered (Reitsema & McIlvaine, 2014; Kyle et al., 2020).

The final step offers researchers the option to export results in a suitable publication format (e.g., CSV, PDF, HTML, DOC, PNG). Researchers are required to input requisite values as stipulated in the author guidelines. Initial tool iterations incorporate typical requirements such as DPI values, table/picture dimensions, and file formats. Future iterations are planned to accommodate additional specifications as necessitated by relevant author guidelines. A preliminary draft of R script is provided in the following github repository as a partial sample of the aforementioned pipeline.

**Discussion**

The study is still ongoing as part of the corresponding author's PhD studies and the discussed applications, codes and results are to be expected. The need to integrate archaeological data such as grave findings and burial datings in future studies ordained the study -from its initial stages already- to set an open framework for collecting data. Therefore, this methodology is considered adaptable to further implementations of research tools and integrations with “Little Minions in Archaeology” as introduced by the CAA community in 2018.

The use of more powerful database tools –such as MySQL, PostgreSQL, SqLite, MicrosoftSQL Server and Oracle Database- is also worth bringing into discussion as it can effectively enhance the methodology of storing material and recording data. However, these tools require a more sophisticated training for anthropologists and therefore can intimidate them before even starting to build their databases. On the other hand, the use of R programming language is considered to be the most appropriate tool for this line of work as it is easier to learn and can handle the amount of data that anthropologists are usually called to deal with (Adler, 2010).

In conclusion, we hail the opportunity to advance bioarchaeological studies through simulation models equivalent to the ones that have been presented in archaeological literature in the last decade (Chliaoutakis 2020). The opportunity is at hand. Human osteoarchaeological material can become a valuable source of input data for simulation models only if bioarchaeologists decide to add programming skills to their scientific quiver.

**Acknowledgements**

The authors express their acknowledgements to all the collaborators participating as members of the research consortium APOIKIA (<https://www.apoikia.org/>). All of them equally created a synergistic and productive environment that is all the more needed in the Humanities. Sincere appreciation is due to the reviewing committee provided by PCI Archaeology. Recommender Christianne Fernee and two anonymous reviewers have offered valuable insights during the preprint evaluation phase.

**Conflict of interest disclosure**

The authors declare that they comply with the PCI rule of having no financial conflicts of interest in relation to the content of the article.

**Funding**

We acknowledge support of this work by the project “ΑΠΟΙΚΙΑ: Ancient DNA analysis in novel multidisciplinary approach of ancient Corinthian colonisation. Ancient Amvrakia and Ancient Tenea as demonstration examples.” (MIS 5056266) which is implemented under the Special Action “Open Innovation In Culture” funded by the Operational Programme “Competitiveness, Entrepreneurship and Innovation 2014-2020 (EPAnEK)” implemented under Regional Operational Programs of the Partnership and Cooperation Agreement (NSRF 2014-2020) and co-financed by Greece and the European Union (European Regional Development Fund).

**References**

Adams BJ, Konigsberg LW (2008) How Many People? Determining the Number of Individuals Represented by Commingled Human Remains. In: *Recovery, Analysis, and Identification of Commingled Human Remains* , pp. 241–255. Humana Press. https://doi.org/10.1007/978-1-59745-316-5\_12

Adler J (2010) *R in a nutshell*. O’Reilly, Sebastopol, CA.

Baxter M (2012) *R-chaeology*.

Buikstra JE, Ubelaker DH (1994) *Standards for Data Collection from Human Skeletal Remains: Proceedings of a Seminar at the Field Museum of Natural History, Organized by Jonathan Haas* (J Haas, Field Museum of Natural History (last), Eds,). Arkansas Archeological Survey.

Calce SE, Kurki HK, Weston DA, Gould L (2017) Principal component analysis in the evaluation of osteoarthritis. *American Journal of Physical Anthropology*, **162**, 476–490. https://doi.org/10.1002/ajpa.23130

Caruso A, Karligkioti A, Selempa G, Nikita E (2023) STARC OSTEOARCH: An open access resource for recording and sharing human osteoarchaeological data. *International Journal of Osteoarchaeology*, **33**, 973–975. https://doi.org/10.1002/oa.3256

Caruso A, Nikita E (2024) Dental diseases and dental wear as a proxy for dietary patterns in Hellenistic-early Roman Menainon, Sicily. *International Journal of Paleopathology*, **44**, 10–19. https://doi.org/10.1016/j.ijpp.2023.11.002

Colleter R, Romain J-B, Barreau J-B (2020) HumanOS: an open source nomadic software database for physical anthropology and archaeology. *Virtual Archaeology Review*, **11**, 94. https://doi.org/10.4995/var.2020.13422

Constantinou C, Nikita E (2022) SexEst: An open access web application for metric skeletal sex estimation. *International Journal of Osteoarchaeology*, **32**, 832–844. https://doi.org/10.1002/oa.3109

Georgiadou A, Ganiatsou E, Xanthopoulos K, Papageorgopoulou C (2023) Estimating the human diet in the ancient Greek colony of Ambracia using Stable Isotopes (δ15N, δ13C) and Bayesian Mixing model. In: *Isotopes and Archaeological Theories of Complexity* , p. . European Association of Archaeologists, Belfast, Northern Ireland.

Hens SM, Godde K (2020) New Approaches to Age Estimation Using Palatal Suture Fusion. *Journal of Forensic Sciences*, **65**, 1406–1415. https://doi.org/10.1111/1556-4029.14485

Hillson S (2005) *Teeth*. Cambridge University Press, Cambridge; New York.

Kubehl K, Temple DH (2020) Reproductive life histories influence cariogenesis: Exploring sex-specific variation in dental caries and survivorship in the human past. *American Journal of Physical Anthropology*, **172**, 376–385. https://doi.org/10.1002/ajpa.24034

Kyle B, Shehi E, Koçi M, Reitsema LJ (2020) Bioarchaeological reconstruction of physiological stress during social transition in Albania. *International Journal of Paleopathology*, **30**, 118–129. https://doi.org/10.1016/j.ijpp.2020.06.003

Lewis ME (2007) *The Bioarchaeology of Children: Perspectives from Biological and Forensic Anthropology*. Cambridge University Press.

Marwick B (2017) Computational Reproducibility in Archaeological Research: Basic Principles and a Case Study of Their Implementation. *Journal of Archaeological Method and Theory*, **24**, 424–450. https://doi.org/10.1007/s10816-015-9272-9

Nikita E (2017) *Osteoarchaeology: a guide to the macroscopic study of human skeletal remains*. Elsevier/Academic Press, Amsterdam.

Nikita E (2020) *INTRODUCTION TO STATISTICS USING R*.

Nikita E, Karligkioti A (2020) *BASIC GUIDELINES FOR THE EXCAVATION AND STUDY OF HUMAN SKELETAL REMAINS*. Zenodo. https://doi.org/10.5281/zenodo.4641953

Nikita E, Mattingly D, Lahr MM (2013) Methodological considerations in the statistical analysis of degenerative joint and disc disease. *International Journal of Paleopathology*, **3**, 105–112. https://doi.org/10.1016/j.ijpp.2013.03.002

Ossenberg N (2013) Brief communication: Cranial nonmetric trait database on the internet. *Am. J. Phys. Anthropol.*, n/a-n/a. https://doi.org/10.1002/ajpa.22377

Reitsema LJ, McIlvaine BK (2014) Reconciling “stress” and “health” in physical anthropology: What can bioarchaeologists learn from the other subdisciplines? *American Journal of Physical Anthropology*, **155**, 181–185. https://doi.org/10.1002/ajpa.22596

ropensci/writexl (2024)

RStudio Team (2020) *RStudio: Integrated Development Environment for R*. RStudio, PBC., Boston, MA.

rstudio/shiny (2022)

Schaefer M, Scheuer L, Black SM (2009) *Juvenile osteology: a laboratory and field manual*. Elsevier/Academic Press, Amsterdam ; Burlington, MA.

Scheuer L, Black SM (2000) *Developmental juvenile osteology*. Academic Press, San Diego, CA.

Schwartz JH (1995) *Skeleton keys: an introduction to human skeletal morphology, development, and analysis*. Oxford University Press, New York.

Seguchi N, Dudzik B (Eds.) (2019) *3d data acquisition for bioarchaeology, forensic anthropology, and archaeology*. Elsevier, San Deigo.

Sjoberg DD, Whiting K, Curry M, Lavery JA, Larmarange J (2021) Reproducible Summary Tables with the gtsummary Package. *The R Journal*, **13**, 570–580. https://doi.org/10.32614/RJ-2021-053

Sokal RR, Rohlf FJ (2009) *Introduction to biostatistics*. Dover Publications, Mineola, N.Y.

tidyverse/readxl (2024)

Vark GN, Howells WW (1984) *Multivariate Statistical Methods in Physical Anthropology: a Review of Recent Advances and Current Developments*. Springer Netherlands, Dordrecht.

Viciano J, López‐Lázaro S, Alemán I (2013) Sex estimation based on deciduous and permanent dentition in a contemporary spanish population. *American Journal of Physical Anthropology*, **152**, 31–43. https://doi.org/10.1002/ajpa.22324

White W (2007) Databases. In: *Advances in Human Palaeopathology* (eds Pinhasi R, Mays S), pp. 177–188. Wiley. https://doi.org/10.1002/9780470724187.ch9

White TD, Black MT, Folkens PA (2012) *Human osteology*. Academic Press, San Diego, Calif.

Wickham H (2009) *ggplot2*. Springer New York, New York, NY. https://doi.org/10.1007/978-0-387-98141-3

Wickham H, Averick M, Bryan J, Chang W, McGowan L, François R, Grolemund G, Hayes A, Henry L, Hester J, Kuhn M, Pedersen T, Miller E, Bache S, Müller K, Ooms J, Robinson D, Seidel D, Spinu V, Takahashi K, Vaughan D, Wilke C, Woo K, Yutani H (2019) Welcome to the Tidyverse. *Journal of Open Source Software*, **4**, 1686. https://doi.org/10.21105/joss.01686

Ελληνική Κυβέρνηση (2015) *Εθνική Στρατηγική Έξυπνης Εξειδικευσης*.

Παπαδοπούλου Β, Βάσιος Α, Παπακώστα Δ, Ξανθόπουλος Κ (2021) ΑΠΟΙΚΙΑ Ένα ερευνητικό πρόγραμμα ανασύστασης της πληθυσμιακής εικόνας της Αμβρακίας την εποχή του κορινθιακού αποικισμού. *Θέματα Αρχαιολογίας*, **5**, 112–123.

1. a short but comprehensive list of the relevant packages includes *readxl* (“tidyverse/readxl,” 2024), *tidyverse*(Wickham et al., 2019), *writexl* (“ropensci/writexl,” 2024), *gtsummary* (Sjoberg et al., 2021) , *ggplot2* (Wickham, 2009), *shiny* (“rstudio/shiny,” 2022). [↑](#footnote-ref-2)
2. for more detailed bibliography on the relevant fields of bioarchaeological analysis see: **age calculation** (AlQahtani, Hector, and Liversidge 2010; Buckberry and Chamberlain 2002; Kemkes-Grottenthaler 2002; Papageorgopoulou et al. 2015b; Schaefer, Scheuer, and Black 2009), **sex determination** (Brůžek et al. 2017; Klales, Ousley, and Vollner 2012; Papageorgopoulou et al. 2015b; Passalacqua, Zhang, and Pierce 2013; Rennie and Ohman 2014; Viciano, López‐Lázaro, and Alemán 2013; Walker 2008; Walrath, Turner, and Bruzek 2004), **height estimation**(Papageorgopoulou et al. 2015a; Raxter, Auerbach, and Ruff 2006; Raxter and Ruff 2018), **dental pathology** (Beck and Smith 2019; Boldsen 2007; D’Anastasio et al. 2013; Garcin et al. 2010; Keenleyside 2008; Primeau, Homøe, and Lynnerup 2018; H. Rathmann, Semerari, and Harvati 2017; Hannes Rathmann et al. 2019; E. C. Scott 1979; Šlaus, Pećina-Šlaus, and Brkić 2004; Smith 1984) and **isotope data** (Reitsema et al. 2022). [↑](#footnote-ref-3)