Age Estimation with Decision Trees: Testing the Relevance of 94 Aging Indicators on the William M. Bass Donated Collection

A Dissertation Presented for the
Doctor of Philosophy
Degree
The University of Tennessee, Knoxville

Kevin Benjamin Dominic Hufnagl
August 2015

Acknowledgements

Ich bedanke mich bei meinen Eltern, meinen Schwestern, und meinen Grosseltern für all ihre Geduld und Unterstützung über die Jahre.

Furthermore I would like to express thanks to the members of my committee for their suggestions and comments that allowed me to polish my dissertation. Thank you to my friends and colleagues at the University for entertaining me otherwise with stories, outings, dinners, and opinions. Thank you for sharing your knowledge and information with me. Thank you to Donna, Heli, and Suzanne for providing me the opportunity to pick their brains, bounce ideas off them, and distracting me when I needed it. They have truly been the cornerstones of my life in Knoxville.

Abstract

Anthropologists have been estimating ages-at-death of skeletons for a long time. A variety of different age indicators has been studied and age estimation methods have been developed in an attempt to standardize the process. Even with all the work that has gone into developing age estimation methods, age estimation of mature skeletons is still very imprecise. This research investigates various age indicator definitions and their performance on an elderly skeletal sample. Using 176 individuals from the William M. Bass Donated Collection, curated in the Department of Anthropology at the University of Tennessee, Knoxville, data were collected on age indicators gathered from fifteen age estimation methods. Ninety-four variables were tested with various decision trees to show patterns among the variables. Regression equations were built using the same variables as the decision trees, and the performance between the two methodologies were compared. The decision trees performed slightly better, with a mean absolute error of prediction of around five years. Variable occurrence was tabulated across various decision tree models. The most common variables are pit shape of the sternal rib end morphology and the phase of the auricular phase. These two variables, along with others commonly selected, present best candidates for building an age estimation method that pertains to older populations.

Table of Contents

Chapter 1: Introduction 1

Chapter 2: Literature Review 11

History of Age Determination 11

Origins 11

Phase Methods 13

Component Methods 21

Multi-Trait Methods 34

Problems in Age Determination 39

Variation in Aging 44

Interpersonal Variation 45

Population Differences 45

Pubic Symphysis 47

Auricular Surface 48

Sternal Rib Ends 49

Cranial Sutures 49

Sex Differences 50

Pubic Symphysis 52

Auricular Surface 54

Sternal Rib Ends 55

Cranial Sutures 55

Intrapersonal Variation 56
Observer Error 59
Interobserver Error 62
Intraobserver Error 64
Method Validation and Reliability Testing 64
Sample Appropriateness 66
Inaccurate Aging of the Elderly 69
Regression to the Mean 71
Age Mimicry 73
Examples of Method Validation 76
Chapter 3: Materials 83
The Skeletons 83
The Sample 84
The Variables 88
Chapter 4: Methods 91
Variable Manipulation 91
Missing Data 91
Additional Variables 92
Decision Trees 93
Analysis 96
Chapter 5: Results 104
Notation 104
Chi-Squared Automatic Interaction Detection (CHAID) Decision Trees 105

Classification and Regression Tree (CRT) Decision Trees 125
Decision Trees using Variable Subsets 136
Comparative Regressive Analyses 146
Results Summary 152
Jackknifed Decision Trees: CHAID 156
Jackknifed Decision Trees: CRT 161
Logistic Regression 165
Chapter 6: Discussion 170
The Use of Decision Trees 170
The Variables 179
Age Distributions 180
Age Indicator Variables 183
Jackknife Decision Trees and Logistic Regressions 185
Specific Results 190
Future Directions 195
Chapter 7: Concluding Remarks 198
References Cited 200
Appendices 208
Appendix A: Variable Definitions 209
Appendix B: Age Method Definitions 214
Pubic Symphysis 215
Auricular Surface 227
Rib Morphology 234

Cranial Sutures.....238

Appendix C: Supplemental statistical output 239

Vita 260

List of Tables

Table 3.1: Demographic distribution of individuals in the study sample 86

Table 3.2: Basic statistics of the age distribution of the study sample 86

Table 3.3: Distribution of Individuals in the study sample based on age groups 87

Table 4.1: Breakdown of Age Groups 102

Table 5.1: Performance Summary Values for various CHAID decision trees 125

Table 5.2: Performance Summary Values for various CRT decision trees 135

Table 5.3: Performance Summary Values for variable subset decision trees 145

Table 5.4: Performance Summary Values for decision tree and regression models 147

Table 5.5: Most common variables used in distinct decision trees and regression models 153

Table 5.6: Breakdown of variable occurrence by aging site 155

Table 5.7: Age ranges and mean absolute error values for jackknife CHAID decision trees 160

Table 5.8: Age ranges and mean absolute error values for jackknife CRT decision trees 165

Table 5.9: Performance Summaries of Logistic Regression and Decision Trees using Age Categories 166

Table 5.10: List of Variables selected for Logistic Regressions and Categorical Decision Trees 168

Table 6.1: Eight most common variables used in distinct decision trees and regression models 193

Table 6.2: Breakdown of variable occurrence by aging site 195

Table A.1: Pubic symphysis variables with references and coding 210

Table A.2: Auricular surface variables with references and coding 211

Table A.3: Rib morphology variables with references and coding 212

Table A.4: Cranial suture variables with references and coding 212

Table A.5: Redefined variables with their origins 213

Table A.6: Variables created out of combining two single sex method 213

List of Figures

Figure 3.1: Histogram of the age distribution of the study sample.....87

Figure 4.1: Basic decision tree with example decisions 94

Figure 5.1: CHAID 100/50 tree with all variables, observations, and all default settings 106

Figure 5.2: CHAID 10/5 with all variables, observations, and otherwise default settings.....108

Figure 5.3: Predicted ages of CHAID 10/5 with all observations plotted against actual age.....111

Figure 5.4: CHAID 30/10 with all variables and observations.....113

Figure 5.5: Predicted ages of CHAID 30/10 with all observations plotted against actual ages 115

Figure 5.6: Predicted ages of CHAID 30/10 without observation 146 plotted against actual ages 117

Figure 5.7: CHAID 30/10 without observations 10 and 146.....118

Figure 5.8: CHAID 30/10 without observations 10, 139, and 146.....119

Figure 5.9: Residuals of CHAID 30/10 without observations 10, 139, and 146 plotted against actual age 122

Figure 5.10: Absolute residuals of CHAID 30/10 without observations 10, 139, and 146 plotted against actual age.....123

Figure 5.11: Residuals of CHAID 30/10 without observations 10, 139, and 146 plotted against predicted age 123

Figure 5.12: Absolute residuals of CHAID 30/10 without observations 10, 139, and 146 plotted against predicted age 124

Figure 5.13: Predicted ages of CHAID 30/10 without observations 10, 139, and 146 plotted against actual ages 124
Figure 5.14: CRT 10/5 decision tree with all variables and all observations 127
Figure 5.15: Predicted ages of CRT 10/5 plotted against actual ages 129
Figure 5.16: CRT 30/10 with all variables and all observations 132
Figure 5.17: Predicted ages of CRT 30/10 with all variables and observations plotted against actual ages 133
Figure 5.18: CRT 30/10 with all variables and excluding observation 10 134
Figure 5.19: CHAID 30/10 with all observations and phase-only variables 137
Figure 5.20: CRT 30/10 with all observations and phase-only variables 138
Figure 5.21: CHAID 30/10 with all observations and component-only variables 142
Figure 5.22: CRT 30/10 with all observations and component-only variables 143
Figure 5.23: Predicted ages for the 13 variable stepwise regression plotted against actual ages 151
Figure 6.1: Histogram of the age distribution of the study sample 181
Figure C.1: CHAID 10/5 with all variables excluding observations 81 and 116 240
Figure C.2: CHAID 20/10 with all variables and observations 241
Figure C.3: CHAID 30/5 with all variables and observations 242
Figure C.4: CHAID 40/10 with all variables and observations 243
Figure C.5: CRT 10/5 with all variables and without observation 15 244
Figure C.6: CRT 50/10 with all variables and observations 245
Figure C.7: CHAID 30/10 with all observations and only author's standard variables 246
Figure C.8: CRT 30/10 with all observations and only author's standard variables 247

Figure C.9: CHAID 30/10 decision tree without age group 1 (23-35) 248
Figure C.10: CHAID 30/10 decision tree without age group 2 (36-40) 249
Figure C.11: CHAID 30/10 decision tree without age group 3 (41-45) 250
Figure C.12: CHAID 30/10 decision tree without age group 4 (46-50) 251
Figure C.13: CHAID 30/10 decision tree without age group 5 (51-55) 252
Figure C.14: CHAID 30/10 decision tree without age group 6 (56-60) 253
Figure C.15: CRT 30/10 decision tree without age group 1 (23-35) 254
Figure C.16: CRT 30/8 decision tree without age group 2 (36-40) 255
Figure C.17: CRT 30/10 decision tree without age group 3 (41-45) 256
Figure C.18: CRT 30/10 decision tree without age group 4 (46-50) 257
Figure C.19: CRT 30/10 decision tree without age group 5 (51-55) 258
Figure C.20: CRT 30/10 decision tree without age group 6 (56-60) 259

Chapter 1: Introduction

Since the 19th century, anthropologists have been observing differences in the skeletons of individuals of different ages (Dwight 1890; Welcker 1862; Welcker 1866). Initially studies regarding progressive changes to the skeleton as people advanced in age were mostly descriptive and focused more on *where* and *how* people's skeletons changed with age, than *when* they changed. Early observations looked at the formation of the skeleton from childhood, the fusing of the long bones, and the formation of the cranium (Ashley-Montagu 1938). Todd (1920) created one of the first systematic approaches to assigning actual ages to skeletons based on how they looked. He observed differences in the pubic bone of a large group of specimens and ascribed these differences to be age changes. Once the differences were understood to be associated with age, Todd managed to separate the bones he studied into 10 groups, and using the actual ages of the individuals in each group, developed a system for estimating the age of an individual based on the characteristics in the morphology of the pubic symphysis.

Since then, anthropologists have made quite a few advances in the realm of age estimation. More aging methods were developed that introduced cranial suture closures, sternal rib end morphologies, and auricular joint changes as viable age estimation indicators. Over the years even more methods for age determination were produced, some of which became staples of the professions, while others never really became a part of mainstream anthropology. Why do anthropologists spend so much time trying to find a good way to determine a person's age-at-death?

Age estimations are a vital part in two major anthropological subdisciplines. In forensic anthropology, age estimations play a key role in the identification of skeletal remains. As part of the biological profiles created by biological anthropologists, the estimated age-at-death of

unidentified remains helps law enforcement officials to identify missing persons or victims of violent crimes. Having an age estimate is helpful because it is much more likely that investigators would be able to identify a missing person if they knew that they had the remains of a "white, male, age 30-35" instead of just "adult white male."

In addition to forensic anthropology, bioarchaeology relies heavily on age estimation in paleodemography. In paleodemography less emphasis is placed on the age estimate of a single individual. Concerned with questions about mortality, morbidity, health, and nutrition, paleodemographers utilize age estimates to get an idea about demographic characteristics of a past population. Many of these demographic measures rely on the age distribution of the population. For instance, in order to establish a mortality profile for a population, demographers need to know how old people were when they died. This information allows calculations of the frequency of death at given ages, which leads to mortality rates, life expectancy, and other mortality-related parameters. These in turn are used to draw inferences about still other topics of past peoples, such as nutrition and health. Now that we know why we want to be able to estimate the ages at which people died, how do we go about getting an age estimate for a skeleton?

Generally, upon receiving skeletal elements for which age estimation is required, an anthropologist will look at a series of different locations on different bones known as age indicators. For most methods, anthropologists then attempt to match their observations at the age indicator site to descriptions or pictures in the literature that describe the various stages that have been identified for that particular age indicator. This is repeated for any other age indicators that are available to be scored. Once each age indicator has been assigned a score, the aging method provides a procedure to convert the age indicator scores to an age estimate or a probable age range for the individual.

Different methods have different ways of treating the age indicator and how to get from the indicator to the age estimate. Some methods look at the whole age indicator and describe different stages as a whole. These are called phase methods, which give one score per method. Other methods, called component methods, split the age indicator up into different characteristics and score each one of these separately, thereby creating multiple scores for a single skeletal element, with each score relating to a different attribute. To get from the score or scores to an age estimate different methods provide different means. Some use a mathematical equation, others a simple reference chart. Still others use a combination of the two or something even more inventive. No matter how the methods work, they all provide the anthropologist with an age range or age estimate for the skeletal remains.

Unfortunately, just because we get an age estimate does not mean it is useful. Forensic anthropology and paleodemography are both interested in getting age estimates for skeletons, but have different priorities and place different values on the accuracy and precision of the age estimates. Accuracy refers to the certainty that the true age of the individual in question is actually in the estimated age range. Precision, on the other hand, refers to the narrowness or broadness of the range, where a narrow range has high precision. As an example, consider someone shooting at a target. If all the shots hit near the center of the target, the shooter would be considered accurate and precise. If the shooter always missed the target to the right but all the shots go to the same location, he would be considered precise but not accurate. Lastly, if the shots are widely scattered but centered around the bullseye, the shooter would be accurate, but not precise. In age estimation, ideally an age estimate should be accurate and precise, but often anthropologists have to balance the accuracy and the precision of their age estimates against one another depending on their needs.

In forensic anthropology, it is up to the analyst to provide a useful age interval to law enforcement officials. In this case "useful" refers to an age estimate that is accurate as well as precise, with emphasis on accuracy. If an age interval is inaccurate, law enforcement will be unable to find the identity of the individuals. If they are given an inaccurate age range of 30-35 years and the person is actually 27, law enforcement would have no reason to include any missing 27-year olds in their search parameters. Obviously, a narrower age range will be more helpful in making an identification, but accuracy is more important. Anthropologists can increase the estimated age interval, thereby sacrificing precision of the estimate, but increasing the accuracy. Besides providing age intervals to law enforcement to facilitate identification, forensics further uses age estimation to establish distributions of age indicators. These distributions allow anthropologists to calculate the probabilities that a person of a particular age has a particular age indicator stage or set of stages. Such calculations provide validity to identifications based on age estimates.

In paleodemography, the interest lies less on the individual and more on the group as a whole. Individual accurate estimates are of limited use when considering a whole group of people. As in forensic anthropology, a precise and accurate age estimate is useful, but when considering the age estimates of a whole group, imprecise estimates are not as problematic. In forensic identification where you only have a single individual, precision helps a lot in the identification. In paleodemography, when looking at a group of skeletons individual imprecise age estimates tend to get drowned out by the rest of the information about the group. Therefore, in paleodemography the emphasis falls on bias—consistent over- or underestimation—instead of precision. An estimation method that always underages skeletons by 5 years is much more detrimental to the field than a method that is 10 years over one time and 10 years under the next,

because those two estimates will average out over the whole population. However, if all the ages are underestimated by 5 years, no amount of data will end up averaging out the bias in the method.

Age estimation appears straightforward enough but it faces several problems. The main issue lies in the age related changes that have been defined. Age-related changes to the skeleton have been extensively studied and defined in children (Scheuer and Black 2000). In younger individuals, the changes to the skeleton and the dentition are very clearly defined and happen in such small time periods that they lend themselves very easily to age estimation. Age indicators for subadults provide a high degree of accuracy and precision. The development of the teeth is so precise that it is even possible to study the effects of developmental and environmental factors on the speed of aging. For instance, it has been shown that low socioeconomic status can have a slowing effect on the maturation of children, their dentition, and their skeletons (Garn et al. 1973a; Garn et al. 1973b).

However, these age-related changes happen rather quickly and once individuals enter their twenties, the age-related changes lose quality. As an individual matures and advances in age, the age-related changes appear less frequently and are much less clearly defined. Changes are still observable in the skeleton, however the associated time frame is no longer precise and a particular change in the skeleton has the potential of occurring over a period of multiple years. The skeleton also shows prolonged periods of static expression where the age indicators do not change, so that an individual looks the same skeletally for a long period of time.

Age indicators also show a high level of heteroscedasticity. Because the speed of the age related changes in the skeleton differ between people, some skeletons appear older than others at the same age. This leads to a much wider range of expression of age indicators in older

individuals than younger ones. For instance, 57-year-old individuals can have skeletons with age indicators that make the skeleton look significantly older or younger than the true age of 57. By comparison, most skeletons of 30-year-old individuals will have age indicators that will not vary as much from the true age.

Further difficulties arise from human variation, which tells us that people are different. As such it should be expected that not everyone ages the same way, or at least not at the same rate. However, due to the scarcity of precise age indicators for adults, it is virtually impossible to consider human variation in age estimation. It becomes another factor that decreases the precision of age estimation. In conjunction with human variation, age estimation is further complicated by the type of data anthropologists are working with. They are not able to observe the changes a skeleton goes through with age as it actually ages. Instead, anthropologists are limited to observing a skeleton only once—in the condition it was in at the time of death. So when anthropologists study age-related changes they get to look at different people who died at different points in their lives. Based on the differences they are able to observe between these individuals, anthropologists infer which changes happen to the skeleton as a person ages. While they cannot know which changes a single individual has gone through, by assuming that all humans go through the same sequence of changes it is possible to outline the general pattern of skeletal aging.

In the fields of forensic anthropology and paleodemography survivability of the remains can be an issue. Dogs and rodents have been known to chew on skeletons (Haglund 1997a; Haglund 1997b; Haglund et al. 1988; Haglund et al. 1989; Klippel and Synstelien 2007) and prolonged burial will also destroy parts of a skeleton. If the age indicator is missing, no age estimation based on that indicator can be performed, reducing the number of sites that can be

studied and thus the overall strength of any age estimation. Even if no traditional age indicators are present, enough research has been done to allow classification of individuals into broad age classifications (e.g., subadult, adult, mature, elderly, etc.). If a portion of a skeletal element is missing, it is possible to misinterpret the portion that is present as the missing portion may be important in age assessment.

However, misinterpretation is not just a problem when bones are poorly preserved. It can even occur with completely perfect osteological specimens. Observer error has been recorded in various anthropological studies (e.g., Baccino et al. 1999; Saunders et al. 1992). This phenomenon occurs when two individuals evaluate the same age indicator on the same individual and arrive at a different answer. It can also happen when the same scientists collect the same data at two different times and find that the data is different. This indicates that the observers changed their approach to the data collection at some point. Most of the age estimation methods used rely on visual identifications based on descriptions. Many of the problems of observer error likely result from the definitions of the age-related changes. In order to evaluate a skeletal element, anthropologists have to interpret what they read as well as what they perceive. Many of the traditional age estimation methods are very general with their definitions and do not clearly differentiate between consecutive age stages. Since the aging process is seen as continuous changes to the skeleton, it makes sense to present consecutive stages in sequential stages that show some overlap. However, this makes it more difficult for observers to know what they should be looking for. With the need to interpret both the written description and the physical expression on the skeletal element, it is not surprising that different anthropologists can interpret observations differently and potentially reach different conclusions, thereby causing observer error.

The last step in the aging process takes all the different age ranges garnered from the different age indicators and combines them into a single age range. If a method uses multiple age indicator observations it generally provides a method for combining them to get an age estimation (Buckberry and Chamberlain 2002; Igarashi et al. 2005; McKern and Stewart 1957; Snow 1983). A few other studies have examined ways of combining age ranges established by different aging methods (Acsádi and Nemeskéri 1970; Bedford et al. 1993; Lovejoy et al. 1985a; Saunders et al. 1992). The simplest way to combine various age intervals into a single age estimate is to look at all of them and eyeball which age range or estimate appears to be the most likely to be correct.

Combining multiple age estimates into a single one becomes easier with experience. As an anthropologist works with more and more skeletal remains, they become more familiar with the methods, which leads to more consistent age assessments. With experience, anthropologists also manage to incorporate additional information into an age estimate, even if the information is not part of the methodology. However, if an anthropologist lacks years of experience working with skeletal remains, they are limited in their age estimation by the methods they employ.

There are many problems to age estimation even if the actual process is very simple. All of these problems combine and stand against the establishment of precise and accurate age estimates.

The purpose of this study is to identify which age indicators and age indicator variables contain valuable information with respect to age. Suppose that the lack of precise age estimation is not due to anthropologists having insufficient information on age related changes to the skeleton or that the anthropologist is applying the methodology incorrectly. The reason for the

lackluster age estimates achieved with all the different methods may not be due to the methodology itself, but rather age indicators that do not actually provide helpful information about the age of a skeleton. This research, rather than look at the performance of age estimation methods on another validation sample, looks beyond the age determination methods and focuses instead on the variables used in the methods. Using decision trees, this study evaluates 94 different age indicators from 15 different methods on four different skeletal elements, to see which of these age indicators is useful in age estimation. Decision trees are used to show their ability to handle large numbers of variables and highlight patterns of interest between age indicators as well as between age indicators and actual age. The following chapters present the exploratory use of decision trees for age estimation and age indicator testing.

Chapter 2 covers the various age estimation methods used in this study as well as research evaluating the performance of these methods. It further elaborates on difficulties that face anthropologists in the area of age estimation. Chapter 3 covers the materials used in this study. It provides a short description of the collection sampled for this research and further outlines the characteristics of the sample with respect to age, ancestry, and sex. Lastly, Chapter 3 covers the data that was collected on the sample, listing all the variables and methods that were used. Chapter 4 outlines the steps performed during the analysis, ranging from variable manipulation to which analyses were performed. It also provides an example of what decision trees are and how the information contained in them is displayed and read. Chapter 5 presents the results of the analyses done. It presents the various decision trees that were constructed, touching briefly on points of interest. It continues to present various regression analyses that were done for comparison purposes. Chapter 6 discusses the implications of this study and the results. It begins by outlining the pros and cons of the use of decision trees for age estimation

with respect to the results of this study. Further, it discusses some general theoretical problems this study faced. The chapter ends with a discussion of the implications of the results and directions future research can take both with respect to age estimation and the use of decision trees. Chapter 7 provides a brief synopsis of the project and outlines its general findings. The appendices provide additional information. Appendix A outlines all the variables that were used as well as their codes for reference. Appendix B reproduces the age indicator definitions from the literature that were utilized for data collection. Appendix C provides additional decision tree diagrams that did not have a good place in the main part of the text.

Chapter 2: Literature Review

History of Age Determination

Origins

The origin of age determination has its beginnings in the 19th century, and as such is also where the age indicator identification process began. The initial scrutiny of age-related changes focused on cranial attributes instead of traits in the postcranial skeleton. In the time when anthropological inquiry focused on cranial typologies and the importance of the skull above all else, most skeletal collections were composed primarily of skulls. Some of these first studies concerned themselves with dental wear (Welcker 1866) and the rate of cranial suture closure (Dwight 1890).

Interest in age determination using postcranial skeletal remains did not appear in the professional literature until the beginnings of the 20th century. Early postcranial interests focused on the age-related metamorphosis of the pubic symphysis. Since then, age estimation research has experienced enormous growth, both in terms of the number of age indicators used and the number of methodologies used to assess them. By virtue of skeletal development studies, nearly every bone of the skeleton can be aged, even if it is only to say that the individual is "mature."

The average adult human skeleton has 206 bones, yet we are born with a much larger number of cartilaginous bone segments (Scheuer and Black 2000). As we grow, these pieces of cartilage ossify and eventually fuse together into what we see in the adult skeleton. The process of skeletal development has been well studied, allowing a relatively narrow age range estimation based on the state of fusion or growth of basically any bone in the human body. By the time we

reach maturity, these same bones have reached their final developmental stage and no longer provide useful information for age estimation, thus making adult age estimation significantly more difficult. In addition, few age-related changes that occur within a narrow time window have been identified in mature skeletons. Most morphological changes in adult age indicators can occur over a wide range of time and thus provide limited information to the age determination process.

Consideration of bone remodeling in age determination is also possible. Signs of osteoporosis or arthritis in the skeleton generally indicate a more advanced age. While there is no direct correlation between a specific age and the development of these pathologies, their presence shifts the age estimate into later years. Besides being a product of aging, arthritis may also develop from heavy or habitual behavior/movement, and osteoporosis can occur early in people due to nutritional deficiencies. Growth of bony spicules and osteophytes in certain regions of the skeleton, such as spinous processes of the vertebral column, can be taken to indicate advanced age. Injury, skeletal trauma, or infection can cause similar bone growths, but these changes will be localized. Only if the extra bone growth is widespread through the skeleton with a fairly even distribution should it be taken as an indicator of advanced age.

After the initial systematic study of age-related metamorphosis of the pubic symphysis, the area of age estimation embraced a more systematic approach to the process. Over the years, the methods used on these age indicators have been expanded, improved, and reinvented (Boldsen et al. 2002; Brooks and Suchey 1990; McKern and Stewart 1957). The previous use of cranial sutures has also been revisited in a more systematic way (e. g., Meindl and Lovejoy 1985). Additional age indicators have been introduced as well, including auricular surface and sternal rib end morphology (İşcan et al. 1984a; Lovejoy et al. 1985b). Although technological

innovations have introduced aging methods on the microscopic and the chemical level (Helfman and Bada 1976; Kerley 1965; Kerley 1969; Ritz et al. 1993) (osteon counting/remodeling and acid racemization), these methods require high-tech equipment or preparation of bone samples before analysis. These kinds of requirements are not very practical for the average researcher.

The focus in the present study remains on easy-to-apply methods where morphological characteristics of age indicators can be readily identified macroscopically. Of all the age indicators that have been studied, four invariably remain as the focus of macroscopic age estimation. These steadfast four include rates of cranial suture closure and morphological changes of the pubic symphysis, auricular surface of the ox coxa, and sternal rib ends.

Phase Methods

In the initial systematic approach to age estimation from the pubic symphysis morphology, Todd (1920) divided the human life span into five year intervals and described the look of the "average" pubic symphysis of a 20 to 25 year old and so on. This method, and methods similar to this describing a series of distinct age indicator morphologies and assigning an age range to each stage, are referred to as "phase methods." Phase methods divide the age continuum into a series of successive stages observable on a particular age indicator, record different characteristics of the age indicator (i.e., degree of billowing and border formation), and observe how these characteristics appear to change with age. Once the basic pattern of age progression is identified, individual characteristics are combined into single paragraph descriptions to represent a single phase. Then, based on the reference sample for each of the stages, a most likely age interval is calculated. All that anthropologists then have to do is look at

the age indicator, identify the particular stage into which it falls, and consult the literature for the age interval corresponding with the identified stage.

In this original approach, Todd (1920) not only recognized the importance of looking at the whole skeleton in order to estimate age, but also acknowledged that most other age indicators (such as epiphyseal fusion) do not help in the aging process beyond the age of 30. On the other hand, the pubic symphysis continues to show change throughout life, even if the changes become less clear after age 40. Todd drew his observations from a sample of 306 white males. He started by describing two 18 year-old skeletons, pointing out features like the symphyseal surface topography, absence of the pubic tubercle and symphyseal margin, and lack of epiphyseal development. He then described a single 19-year-old specimen and subsequently lumped the three individuals into a "first post-adolescent phase" (1920: 301). This phase he described in list form citing the states of four characteristics: surface, nodule, margin, and extremities. He continued in this fashion, first describing a set of specific pubic bones and then lumping them into a phase based on the ages of the specimen. He summarized the phases in short bullet-point fashion, frequently describing them in terms of changes from the previous phase. The age ranges for these phases are between two and four years until age 30, after which the phases generally span five-year intervals. The last phase Todd described is the tenth, which is for those individuals over 50 years of age. He then describes and discusses 13 samples that did not fit correctly into his defined age categories; approximately half were older than their symphyses suggested, and the other half were younger. Todd manages to explain several of these misclassification cases, citing alcoholism, dwarfism, juvenilization, and general sickliness of the individuals. Nevertheless, he does admit that a couple of these skeletons not falling into the correct phases had no discernible reason for their misplacement (Todd 1920: 315-317).

With his ten phases, Todd was able to outline the general age progression of the pubic symphysis quite succinctly. Earliest age indicators are the rugged surface topography of the symphysis. Over the course of the following 15 years, the symphysis sees the completion of the dorsal margin, the ventral rampart, and the upper and lower extremity. After the general form of the pubic outline is complete, the symphysis becomes stable for a bit, then forms bony outgrowths and erodes in the tenth phase (1920). Todd's description of the pubic symphysis age progression is very thorough, although some of his detailed description is lost in his phase summaries. Much of the terminology used by Todd has been adopted in subsequent methodologies pertaining to the aging of the pubic symphysis. Todd believed that his phases were directly related to age, although some variation existed with respect to the speed of age progression.

In 1921, Todd repeated his study of age progression of the pubic symphysis, looking at white females, and males and females of mixed ancestry. He used a sample of 90 individuals to look at the aging pattern expressed by men of a mixed ancestry. His presentation took the same form as in his previous work, discussing individual specimens before grouping them into a shortly summarized phase. In the sample, he found three examples for both age acceleration and delay. He noticed that some of these misclassified skeletons suffered from general skeletal age acceleration or delay where the whole skeleton did not correspond with the chronological age of the individual. Others showed localized age acceleration or retardation where the age of the skeleton did not correspond with the age of the pubic symphysis, in turn failing to match the chronological age. In the end, the ten phases outlined for his admixed population had almost the exact same definitions as those for his previous study.

For the 47 white female pelves, Todd maintained his established pattern, first discussing the appearance of several specimens of similar ages and then developing a summary description of the set of symphyses (1921). Once again, several female pubic symphyses showed deviation from the established phases. In general, the age progression noted by Todd was very similar between the sexes. For his sample of 22 female pelves of mixed ancestry, Todd's description of their pubic symphyses was terse and five were misclassified within his ten age phases. Generally speaking, the phases outlined by Todd for the two female samples were descriptively nearly identical.

Summarizing his four different samples, Todd concluded that the pubic symphyseal metamorphic changes were the same across the samples. Slight differences were present in timing but the descriptions of the phases remained the same. The ten phases spanned three stages: a post-adolescent stage, a second stage in which the pubic symphyseal outline formed, and a third when the symphysis ceased to change and began exhibiting secondary changes. Differences in age progression were more notable between the sex samples than between ancestral groups (Todd 1921).

While the Todd aging method remained the main pubic symphysis phase system utilized by researchers, problems with its accuracy led to a redefinition in 1985 (Meindl et al.). Failing to get satisfactory levels of bias when employing the ten phases outlined by Todd on a sample of 96 skeletons, the authors cut the number of phases in half to increase their accuracy. Their five stages are based on the general morphology of the symphyseal face. Where Todd chose skeletons of a similar age and created phase descriptions based on what he saw in those few samples, Meindl and colleagues considered the whole set of pubic symphyseal changes before breaking it into five progressive age phases. With the approximately corresponding Todd phases given in

parentheses, the Meindl et al. phases were pre-epiphyseal (I-V), active epiphyseal (VI), immediate postepiphyseal (VII), maturing and pregenerative (VIII), and degenerative (IX-X) (Meindl et al. 1985). As shown above, the main difference with the Meindl et al. five-phase system is that their first phase encompasses the entire first half of Todd's aging system. The authors mainly based this phase collapsing on their finding that many of the traits observed by Todd occurred over too wide an age range to allow for a confident claim of a two or three-year period as Todd had done (1920). For each of their five phases the authors wrote descriptions of which features developed during each phase and provided observed age ranges for each feature observed in their sample. Their results showed that individual features have a large range of occurrence, in many instances exceeding seven years. Along with their individual feature frequency they presented a summary statement for each of their phases. More importantly, they reproduced a nice series of 45 photographs of pubic symphyses showing a variety of pubic symphyseal morphological expressions. The main drawback of the system suggested by Meindl and colleagues may be the lack of a summary table or list for ease of application of the method.

Todd's ten phase system remains in use today, while Meindl's simple five phases are rarely mentioned. Likely, this comes from a subsequent reinterpretation of Todd's system by Brooks and Suchey (1990), who modified Todd similarly to Meindl and colleagues, but derived six stages. Based on studies of interobserver error with modern samples, Brooks and Suchey combined Todd's phases 1, 2, and 3; 4 and 5; and 7 and 8 into just three phases, keeping the original phases 6, 9 and 10 of Todd's system. Based on a sample of 739 male and 273 female pubic sets collected at the Los Angeles County Office of the Chief Medical Examiner, the team introduced a separate male and female 6-phase system, calling it the Suchey-Brooks method. Based on their combination, the disapproval of Todd's first five phases already apparent in

Meindl's revamping, is re-expressed. Unlike Meindl however, Brooks and Suchey also present summary descriptions of their phases and produce some photographs to illustrate their phase descriptions (1990). They went a step further, producing casts of the pubic symphysis phases to allow researchers access to three dimensional examples. For each sex, twelve casts are available, two per stage, with one representing an early expression of the stage and the second representing a latter expression of a pubic symphysis in the same stage. These casts, and the ease of application of the Suchey-Brooks methods, are likely reasons why this system is still one of the most commonly used phase systems for the pubic symphysis.

While the pubic symphysis has been studied extensively, it is not the only age indicator to be classified into age-progressive stages. The auricular surface (the joint surface on the os coxa between the sacrum and the innominate) has also been identified as a skeletal site that undergoes generally predictable morphological changes over time. The main phase aging system for the auricular surface was established in 1985 (Lovejoy et al.) Based on a total sample of 764 well-preserved auricular surfaces, this method outlines the changes that the auricular surface appears to undergo with age. Lovejoy and colleagues are advocates of seriation, a process by which they looked at a large number of the same skeletal element and sort them into a long series based on the feature they are investigating.

Lovejoy and his colleagues identified several traits on the auricular surface that showed progressive change with age. Before describing their age stages, they clearly defined areas of the auricular surfaces (e.g., demiface, apex, and retroauricular) and general terms (e.g. grain, billowing, porosity, and density) used in the subsequent age descriptions (Lovejoy et al. 1985: 17-18). The authors looked at the progression with age of seven different features/regions of the surface: grain and density, macroporosity, billowing, striations, apex, retroauricular area, and

transverse organization (Lovejoy et al. 1985: 18-19). Based on these changes, they described eight different age phases, the first six covering the years 20 to 49 in five year intervals and the last two corresponding to the age ranges of 50-60 and greater than 60 years respectively. The authors suggested this method should, in essence, be used in the same manner that Todd's method is employed. Recognizing that not all specimens are likely to fit neatly into their age phases, the authors stated that analysts of the auricular surface should determine the "paramount age criteria that best represent the biological aging process and correspond to the designated phase" (Lovejoy et al. 1985:26). These paramount features are the granularity of the surface and the topographical changes from billowing over striations and densification to erosion and breakdown. Changes to the apex and the retroauricular area are classified as auxiliary aging features used to increase or decrease the age estimations within the particular stage. Unlike some other methods, the authors claim their method can be applied equally to males and females. Additionally, if a female exhibits a preauricular sulcus--a groove adjacent to the border of the auricular surface, this feature should be ignored in the evaluation (Lovejoy et al. 1985). No other major phase methods for age determination of the auricular surface have been established.

İşcan and his colleagues (1984a; 1985) looked at the progressive changes to the ends of the fourth rib where it attaches to the sternum. In this study, the authors established separate age estimation methods for males and females using only white individuals. Observing the state of the sternal rib end on 118 male and 86 female ribs, İşcan et al. developed a nine phase age determination system for each group. They chose to observe the fourth rib not because it was the most age-diagnostic, but rather because it is the most readily available rib in postmortem exams. The initial method was established based on the male sample (İşcan et al. 1984a). Individuals below the age of 17 were found to have no discernible changes to their rib ends and thus became

the initial starting phase for the method. The method identifies three major morphological traits to consider for age estimation: the shape of the pit at the rib end, the edge around the pit, and the overall quality and density of the bone. The rib end, initially slightly billowed, grows flat and then begins indenting with age. The pit grows deeper and deeper starting out with v-shaped walls that eventually widen into a u-shape. The edge around the pit starts out thick, rounded, and wavy and with age first become irregular, and then thin and sharp-edged. The quality of the bone consistently decreases with age, losing both strength and density.

The nine phases—eight if you do not count the initial default pre-17 stage—focus on these traits. The first half of the stages have narrow age intervals, which, as with other phase methods, increase in width with age where later stages tend to have a greater intervals. Unlike most aging methods, however, in Isçan's rib method, the male age range reaches as high as 85 years. Stages 0 through 4 cover the age range up to thirty years, while stages 5 through 8 cover the remaining life span separated into approximately ten year intervals (İşcan et al. 1984a). The method established for the female sample is very similar to the male method, also having nine phases with the first five (0 through 4) covering a much narrower age range (14 to 28) than the second half of the phases (İşcan et al. 1985). Each of the earlier stages has about a three to five year predictive age range, while in the latter four stages of the method the predictive age intervals increase to 10 to 15 years, substantially larger than in the earlier stages.

As seen previously, phase methods attempt to boil down a complex process at a particular skeletal site. Where several events are happening and several features are changing with age congruently, phase methods combine common sets of exhibited traits into general appearance-types. These are then utilized to categorize bones and develop the age determination standards for the particular age indicator.

Component Methods

Component methods of age estimation take a different approach to the various skeletal age indicator sites. Rather than looking at the whole site and identifying a "look" based on a complete description, component methods, as the name suggests, break the age indicator down into individual sections or traits. Examples of individual components would be the degree of ventral rampart completion on the pubic symphysis or the amount of macroporosity on the auricular surface. These components are intended to be individually and independently scored into several descriptive stages. These component scores are combined in a variety of different ways to arrive at a combined score that is tied to a particular age range, just as phases are tied to particular predictive age estimation intervals. A few methods skip the composite score and instead directly use the component scores to calculate an age estimate.

Component methods are slightly more complex in that they tend to utilize more statistical approaches. Initially, however, their main justification was that they appear to allow for variation where the phase methods do not. McKern and Stewart (1957) suggested that Todd's original approach, while doing a good job describing successive age periods, was too static and failed to allow for variation in the age expression of the pubic symphysis. They state that in Todd's symphysis method, as in any other phase method, only those pubic symphyses that are close to the descriptions can actually be aged reliably. While this is not completely true (as most phase methods after Todd address unclear observations), it is nevertheless the case that phase methods do not provide a systematic way to evaluate an age indicator exhibiting age characteristics of multiple phases.

In their extensive work on aging patterns in the skeleton, McKern and Stewart (1957) focused primarily on the timing of epiphyseal fusion throughout the skeleton of Korean War

POWs and soldiers killed in action. They extended their investigation upon study of the innominate and suggested an alternative method to pubic symphysis aging that "takes into account all variations" (1957: 72) rather than defining ten static modal phases. They started with Todd's description of the pubic symphyseal morphology and identified nine characteristics central to Todd's phases. These nine traits are: upper and lower extremity, billowing, dorsal margin and plateau, ventral beveling and rampart, the superior ossific nodule, and the symphyseal rim. Through observation, McKern and Stewart combined these nine characteristics into three regions (dorsal plateau, ventral rampart, and symphyseal rim), for each of which they defined a pre-beginning state 0 and five developmental stages (1957:72-79). In this manner they were able to identify the symphyseal state of each pubic symphysis by a three number series, i.e. XYZ, each of the digits representing the stage of one of the three regional components. McKern and Stewart presented the age ranges for the stages of each component individually first, but then continued to create a "composite score" (1957). This composite score is simply the sum of the three component's stages, so that a symphysis with a ventral rampart of stage 3, a dorsal plateau of stage 3, and a symphyseal rim of stage 4 would have a composite score of 10. The authors calculated predictive age ranges and mean ages for each composite score ranging from 0 to 15.

While this method may seem quite similar to a phase method as described earlier, it is somewhat different. The fundamental difference in this method is that the age indicator is broken up into several different aspects, each of which is analyzed separately, with the method then providing a means for combining the separate components into one age estimate. The argument that this type of method allows analysis of more varied expressions is correct because it is able to estimate age for different combinations of characteristics that are not tied together in a one-to-one basis. Phase methods do allow for age estimation based on age indicators exhibiting

characteristics of various stages, however, these types of estimations are generally not part of the methodology and become available to expert analysts only with extensive practice with the aging methods.

As the McKern and Stewart three-component method was established from casualties of the Korean War, their demography was extremely limited. Their 375 subjects were all males ranging in age from 17 to 50, with only seven individuals in the 40 to 50 year range (1957). As Todd had already established a factor of sexual dimorphism in the process of pubic symphysis aging, a few years later a second three-component method was developed. This time, Gilbert and McKern focused on the aging of the female pubic symphysis. Like McKern and Stewart, Gilbert and McKern start by looking at the phase descriptions in Todd's (1921) work on the female pubic symphyses (1973). They defined three components that were the same ones used by McKern and Stewart in 1957. Furthermore, the six stages defined for each component were extremely similar to the corresponding definitions observed sixteen years earlier on male symphyses. Based on a sample of 103 individuals, this method covered females in the 13 to 57 year age range (Gilbert and McKern 1973). For both of these original pubic symphysis component methods, casts were made for each of the six stages in each component (McKern and Stewart 1957; Gilbert and McKern 1973).

A few years later, Hanihara and Suzuki (1978) took a different approach to the pubic symphysis. Rather than repeating the technique of the three-component methods that lumped nine characteristics into three components based on similarity, the authors scored the purely basic traits individually with no lumping of characteristics. Hanihara and Suzuki identified seven traits and assigned each of them between two and four states. Based on their data from 70 Japanese pubic bones they developed a multiple regression equation (MRA) as well as a

quantification theory model I (QMI). The QMI models are simply regression models that treat the independent variables as discrete numbers and use dummy variables for each age indicator stage. With these equations, the age estimation became a simple computational sum as each trait's score added a certain number of years to the baseline. The difference between the two models is a simple one, but one that is of great importance in the realm of current age estimation techniques. Multiple regression treats each of the variables as continuous while QMI treats each trait like a discrete variable. While the authors found that both of their models were good, QMI was slightly better. Multiple regression tended to overage the young and underage the older individuals; a pattern which was reversed for the QMI model. As the age range for their sample was limited, Hanihara and Suzuki (1978) suggested their method should only be used on individuals between the ages of 18 and 38.

This lack of covered age range led to an improvement to the method by Chen and colleagues (2008). Working with 262 Chinese pubic symphyses, the authors worked off the traits outlined by Hanihara and Suzuki (1978) and the descriptions given by Brooks and Suchey (1990). It appears that they adopted some trait definitions directly from Hanihara and Suzuki (1978), possibly with slight wording changes. Other traits were modified to varying degrees, and a few were created specifically for this method to supplement the others. In the end, they defined nine different traits each with between 3 and 5 stages (2008). Chen and colleagues used three different methods to combine the individual trait scores into an age estimate. Like Hanihara and Suzuki before them, they used multiple regression and QMI. On each of these they used gradual regression analysis to develop four age estimation equations: two MRA and two QMI equations, one each for the regular method and one for the gradual regression version. The gradual regression appears to be a correction of the coefficients of the regression equations.

Chen et al. provide the equation for the gradual regression equation in their work but do not explain it.

Several issues become evident after reading this study. The authors did not validate their method nor did they elucidate the differences between their set of age estimation equations or indicate when to use which one. They did, however, consider that the method addresses some of the shortcomings of the previous methods of Todd (1920; 1921), Brooks and Suchey (1990), and Hanihara and Suzuki (1978). The method outlined by Hanihara and Suzuki was only viable for the relatively narrow age range of 18 to 38 years, while the Chen et al. method worked for individuals between the ages of 14 and 70. Similar to previous arguments (Hanihara and Suzuki 1978; McKern and Stewart 1957), the authors argued that individual component methods are superior to phase methods (i.e., the Todd and Suchey-Brooks methods), because they allow for flexibility in the expression of the age indicator and a large range of age trait combinations (Chen et al. 2008).

As with phase methods, the auricular surface is secondary to the pubic symphysis in the amount of attention it receives in component methods. Buckberry and Chamberlain (2002) approached the auricular surface similarly to the way McKern and Stewart reworked the aging based on the pubic symphysis. Validation studies of Lovejoy's phase system showed tendencies to underage the old and overage the young (Bedford et al. 1993; Murray and Murray 1991; Saunders et al. 1992). The traits used by the Lovejoy method also appear to develop independently of one another. Buckberry and Chamberlain tried to address these issues by switching the system to components. Working off the phase descriptions developed by Lovejoy et al. (1985b), the authors defined five traits: degree of transverse organization, surface texture,

microporosity, macroporosity, and apical changes. Each of these traits has a possibility of three to five different stages (Buckberry and Chamberlain 2002).

Buckberry and Chamberlain used simple summation of the trait scores to aggregate the components. These composite scores range from 5 to 19 and were combined into seven age stages based on similar predictive age estimation ranges of the composite scores. Based on their sample size of 180 individuals, their seven compiled age stages ended up covering the age range of 16 to 92 years, with the median ages for the stages ranging from 17 to 73 years. In order for other anthropologists to use the method they also calculate posterior probabilities based on a uniform Bayesian prior. This analysis provided probabilities of an individual being in a ten-year age range given their auricular component composite score (2002).

A second component method for aging of the auricular surface was published by Igarashi and colleagues (2005). Their method is actually quite unique and unlike any others discussed so far. It is still a component method but instead of defining a range of expressions for each trait and then dividing that range into multiple scores, Igarashi et al. defined traits and only scored their presence or absence, thus creating a binary component aging system. The major benefit of using binary traits is that it should lower inter- and intraobserver errors. This is expected because there are fewer choices for the observer to make and a smaller range of expressions to identify. Igarashi and colleagues identified 13 traits and scored them based on whether or not they were observed on the auricular surfaces. They were divided into three categories with four traits describing the auricular surface relief, five pertaining to the surface texture, and the last four scoring extra bone growth on, around and along the auricular surface. While the relief and texture variables are reminiscent of stages from individual traits from some of the other methods, treating them as individual traits allows for complete freedom in the data recording process and

no choices have to be made when the auricular surface exhibits features of two adjacent stages. With this method the auricular surface either has a feature or it does not. The only difficulty that remains is the correct identification of features.

Working with a set of 700 modern Japanese skeletal remains (262 female and 438 male), Igarashi and colleagues investigated the distribution of their 13 traits. Their male specimens ranged from 16 to 89 years of age while their female sample represented the range from 11 to 96 years. They used the age of 23 as a cutoff between adults and subadults (Igarashi et al. 2005). Initially, they attempted to determine groups of traits based on similar patterns of presence, but as no patterns could be elucidated, the authors classified each of their thirteen traits as either an older or a younger trait based on their observations. This definition was based on the relationship between the average age of those skeletons with and without a particular trait. Five of the thirteen traits were identified as "younger" if present on the auricular surface. Therefore absence of these traits would be categorized as the older state. By summing the number of older features in a particular auricular surface, Igarashi et al. created composite scores for the auricular surface. Based on similar characteristics of some of these scores, the 13 different composite scores were lumped together into four composite groups. Yet, as these groups showed substantial overlap in their probability density functions they were not deemed discriminating enough to be useful in age estimation.

Alternatively, Igarashi et al. created multiple regression equations for each sex and for the combined set of all remains. They developed one equation using all traits and one equation using stepwise regression only selecting those variables with significant age correlations. They found that their regression equations tended to overage the younger individuals and underage the older individuals in their sample, a pattern that is observed quite frequently in the realm of age

estimation (Igarashi et al. 2005). They conclude their analysis with a detailed discussion about how their variables contributed to the age estimation and compared their method's performance to several others. Interestingly, they concluded that no single method is useful for determining age by itself.

Age estimation methods for the ribs are much rarer than those utilizing the innominate. İşcan and colleagues (1984b) outlined a simple component analysis using a sample of 93 fourth right ribs from white men over the age of 17. Working with a sample of predominantly younger individuals—only 25% of them being over the age of 50—Isçan developed a system for age estimation similar to that of McKern and Stewart (1957). The authors broke up the appearance of the sternal rib end into three components: pit depth, pit shape, and rim formation. Each of these components was subdivided into six different states of expression. Although pit depth was actually measured, the researchers grouped the different depths together instead of using the actual measurements in an apparent attempt to keep all three components identical in structure. Their analysis, much like many age estimation studies before them, showed that the recorded morphological changes occurred more rapidly early in age and slowed down in later life, again making the process of differentiating older individuals much more troublesome. They further analyzed a simple composite component sum that ranged from 3 to 15 (although with stages of zero the possible composite score should go from 0 to 15) which indicated a slow, yet steady increase in mean age. The ranges for successive composite scores showed substantial overlap (İşcan et al., 1984b). The most likely culprit of this phenomenon outside the method itself is likely sample size. With a sample of 84 (those of the original 93 having composite scores) and 15 groups, few groups had a large number of individuals. Only four composite scores were attributed to more than seven individuals each. With such small groups, calculations of group

characteristics, such as mean age and age prediction intervals, yield quite uncertain results. Although ribs, especially older ones, generally have low survival rates in a taphonomic context, İşcan and colleagues point out that the age range investigated by their study is much greater than that of the comparable McKern and Stewart pubic symphysis study (1957) and allows age estimation of individuals into their early 60s.

Estimating age based on cranial suture closures may be the oldest means of age estimation (Ashley-Montagu 1938). It is well established that the fusion of epiphyses to diaphyses of nearly every bone in the human body is quite regular, barring environmental and nutritional factors (Garn et al. 1973b). Much like the formation and eruption of teeth, these fusions allow quite narrow age estimations for the first two decades of an individual's life. However, after reaching early adulthood most epiphyses have fused and are no longer able to provide data for age estimation.

Cranial suture union has been observed since the sixteenth century, originally addressed by Vesalius and his pupil, Fallopius (Ashley-Montagu 1938). Early on, the etiology and process of suture closure was believed to be tied to cranial development with suture closure influencing cranial growth itself. It was later understood that the development of the brain is not hindered by the process of suture union. Some of the first examples of researchers attempting to connect age to suture closure appeared in the nineteenth century; however, very early statements by scientists were often general and based on small sample sizes. While many studied the changes to the skull with age (see Bolk 1913; Broca 1861; Dwight 1890; Frédéric 1906; Mazzi 1918; Pommerol 1869; Ribbe 1885; Topinard 1885; Zanolli 1908 in Ashley-Montagu 1938), few looked at suture closure, and none outlined a clear connection between age and suture closure. Most scientists concluded suture closure to be too variable to be of use for age association. Conclusions

compiled from their disparate works include the following (Ashley-Montagu 1938; Todd and Lyon 1924):

1) In 1870, Sauvage concluded that suture closure did not begin until 45
2) In 1890, Dwight found that most sutures begin closure before the age of thirty and that union tended to commence on the endocranial surface of the vault.
3) In 1905, Parsons and Box agreed with the latter statement but suggested that suture obliteration does not take place nominally until after 30 years of age
4) Frédéric, in 1906 concluded, based on primarily ectocranial suture closures of a sample of 374 individuals, that age estimation based on suture union was only good to about a decade and that it came with high degrees of individual variation.

Questioning the reliability of Frédéric's sample and age estimates, Todd and Lyon (1924) undertook a comprehensive look at the progression of endocranial suture closure. Using 267 white males selected from a large collection they had accumulated over several years, the authors looked at nine different cranial sutures. These were broken up into 18 different suture segments such as the superior masto-occipital suture and the pars asterica of the lambdoidal suture. They pioneered scoring each of the suture sections on a numerical five point scale, where 0 represented an open suture and 4 a completely closed one. The integers between represented successive quarters of sutural union. In this way, each age from their sample received an average closure score based on the individual skulls evaluated. This allowed them to say that the average state of closure of the bregmatic part of the coronal suture at age 29 was 3.4, meaning that most skulls of that age were mostly to completely closed. For each suture segment, Todd and Lyon describe the process of suture union as periods of rapid or gradual union or cessation of fusion based on line graphs of age versus average suture union score. Each suture has its own unique

progression but the description take generally the same form. "Suture closure begins around age X, progresses rapidly until age Y to a score of W, and then progresses more gradually until complete closure is achieved at age Z" (e.g., Todd and Lyon 1924: 345). While the exact numbers do not match across sutures, Todd and Lyon found two patterns that repeated throughout the cranium: 1) after suture union commences sometime in the 20s, it progresses rapidly until 30 years of age; and 2) around the age of 30, sutures across the skull enter a period of inhibited fusion. The authors provided nice, easy to understand summaries of the progression of suture closure for each of their 18 suture portions. While these outlines represented a good compilation of information about the closures of endocranial sutures, the authors nevertheless did not present any means to utilize this information to estimate an individual's age based on the state of the sutures in their cranial vault (Todd and Lyon 1924).

In subsequent papers, Todd and Lyon (1925a; 1925b; 1925c) outlined their findings with respect to ectocranial suture closures of the same sample of white individuals they used for the endocranial sutures, as well as the endo- and ectocranial suture closures of a sample of 79 African American crania. For each of their 18 suture segments, both on the outside and the inside of the skull, the authors described the "modal" progression of suture closure. Sometimes they used bullet points and sometimes just a paragraph, but the format mentioned above is maintained. They outlined periods between certain ages of fast or slow suture union. Many other periods were termed "periods of oscillation" when Todd and Lyon could not determine a clear pattern of union based on their line graphs. All in all, they found ectocranial sutures to express the same general pattern of suture closure as their endocranial counterpart. Ectocranial sutures tended to close slower and exhibit a higher degree of individual variability and thus were concluded to be less reliable age indicators than endocranial sutures (Todd and Lyon 1925a).

After Todd and Lyon, few people revisited suture closure. Of those who did, none became advocates of using it for age estimation due to the great variability associated with the process (Brooks 1955; McKern and Stewart 1957; Singer 1953). As part of a larger project, Acsadi and Nemeskeri also looked at cranial suture closure (1970). Looking both endo- and ectocranially, the two authors scored 285 crania on a five point scale. While Todd and Lyon also used a five-point scale, the one employed by Acsadi and Nemeskeri was much more subjective, using definitions like "incipient closure" and "advanced closure" (1970: 115). The authors only looked at ten suture sites along the coronal, sagittal, and lambdoidal sutures of the vault. Their findings correspond to those made by Todd and Lyon nearly fifty years earlier: the ectocranial sutures fuse later and at slower rates than the endocranial sutures. Unlike the earlier study, however, these researchers did not provide the data for the individual suture sites, but rather presented their data on their calculated mean suture closure, a single average of the ten scores per skull. With their mean suture closure they calculated a single second-degree regression equation linking chronological age and mean suture closure (Acsadi and Nemeskeri 1970). The authors concluded that suture closure can provide valuable information for age estimation; however, that they should not be used alone, but in conjunction with other age indicators.

In 1985 Meindl and Lovejoy revisited age determination, this time based on suture closure. They focused on ectocranial sutures both because these have been shown to remain open longer than endocranial sutures (Todd and Lyon 1924), and because information is lacking for suture closure of individuals of advanced age. Observing 17 points on 236 crania, each point along a suture for a one centimeter length was scored from 0 to 3 (from open, less and more than half fused, and completely fused). With this scoring system, the only subjective judgment was between scores of 1 and 2, whether or not the suture was fused more than 50%. Of the 17 points,

the authors rejected seven due to difficulty assessing them, asymmetry, or weak ties to age. Of the remaining 10 sites, lambda and the midlambdoid showed the highest, although still only moderate, correlation with age. Meindl and Lovejoy combined the suture points into two sets: the vault set comprised of seven points and the lateral-anterior set made up of five points. For each set they summed the scores into a single composite score. Both composite scores are helpful in estimating age and show much lower deviations than corresponding endocranial studies. As the sutures of the lateral-anterior group tend to remain open longer, they are more useful for estimating in the upper ages. Results of this study allowed Meindl and Lovejoy (1985) to show that ectocranial suture closure can provide valuable information for age estimation in conjunction with other age indicators. They agreed that high individual variability makes sutures unlikely to be useful as a solo age indicator, but that suture closure, like all other age indicators, should be used in combination with others to arrive at the most accurate age estimates.

In a short study, Mann and colleagues (1987; 1991) evaluated 36 crania for four sutures of the hard palate: the incisive, the anterior median palatine suture, the transverse palatine suture, and the posterior median palatine suture. The researchers actually measured the length of obliteration in the sutures and calculated the percentage of the whole suture length that was obliterated. This percentage was divided into five scores based on incremental 25 percentage point steps (0-4). They only presented cursory data findings and did not perform statistical analyses to create a correlation between palatine suture closure and age. Based on their small sample of 36 individuals, they concluded that the posterior median suture does not begin closure until after 25 years and neither the anterior median and the transverse palatine sutures commence obliteration until 43 years. Above the age of 60, Mann and colleagues stated that two of the four

sutures (the incisive and the posterior median) are generally completely obliterated (Mann et al., 1987).

As can be garnered from the summaries presented above, component methods of age estimation are just as common as phase methods. The component approach could be viewed as a more basic attempt at defining the changes that occur in an age indicator. Rather than looking at an age indicator all at once, component methods focus their attention on individual pieces of an age indicator. Both approaches have merits as well as drawbacks; these will be addressed later in this paper.

Multi-Trait Methods

Component methods discussed above are more complex than phase methods because they have to address the issue of establishing a single age estimate from multiple points of analysis. These points are however all contained within the same age indicator. Interestingly, almost since the very first anatomists attempted to establish ties between changes in the skeleton and the age of the individual, claims have persisted that age estimates based on a single age indicator are useful, but should be used in conjunction with other methods and indicators to provide the best possible age approximations. Yet, until fairly recently, no one addressed how these age indicators should be combined. One would presume the original implication was that the age estimations would be combined by a sort of averaging method or simply based on the experience by anthropologists on previous cases. A few have presented methods of age estimation that utilize more than a single age indicator.

One of the first true multi-trait approaches was introduced by Acsádi and Nemeskéri (1970). They introduced their "complex" method, which combined age estimations based on

cranial suture closure, pubic symphyseal morphology, and radiographic translucency of the proximal aspects of the humerus and femur. Using their own work, Acsádi and Nemeskéri broke both the processes of cranial suture closure and symphyseal age progression up into five stages. Further, they defined six translucency stages each for the humerus and the femur, determined primarily by the extent of the medullary cavity toward the proximal end of the bone, and the degree of rarefaction seen in the spongy bone of the proximal ends. Based on a sample of 105 skeletons, the authors calculated the mean ages and standard deviations for each of the 22 stages. Their "complex" method focused on these means as well as the upper and lower limits of the age estimation intervals. The whole method has three steps. Initially, the pubic symphysis should be used to determine generally if the individual is younger than 50, about 50, or older than 50 years of age. Secondly, as with all other aging methods, the correct age stages are identified for each age indicator available, which provide four age ranges and four mean ages. Lastly, if the initial assessment was that the individual was younger than 50, the lower limits of the age intervals are averaged. Conversely, if the individual looks older than 50, the upper limits are averaged into the final age estimate; if they are around 50, the means of the multiple indicators are averaged together (Acsádi and Nemeskéri 1970). The authors demonstrated how this system works on three individuals, but they also showed that both severe over- and underestimation of age can occur when using this method.

Following introduction of the "complex" method, questions were raised about the validity of simply averaging together age estimates from different age indicators. Many anthropologists thought that giving each indicator equal weight was the incorrect approach. One could argue that some methods are more valid than others and should be more heavily considered in the combining of age estimates. Others could also argue that newer methods—methods not fully

validated with the scientific community, or methods with which the analyst is not as familiar or comfortable—should not be given as much consideration as more established and supported methods. Consequently, the multifactorial method was introduced (Lovejoy et al. 1985a; Meindl et al. 1983). The multifactorial method was developed with five age indicators: auricular surface, pubic symphysis, cranial suture closure, dental wear, and trabecular complexity of the proximal femur. Individuals from the archaeological Libben population were aged based on these five age indicators. An intercorrelation matrix was generated on the basis of the age estimates from the five age indicators and a principal component analysis of the matrix performed. The first component is ascribed to represent chronological age and each age indicator's correlation with the first principal component is taken to be that indicator's weight. These weights are used to calculate the weighted average of the available age indicators for each individual to arrive at the final "summary age" estimate (Lovejoy et al. 1985a).

The whole process may sound quite intimidating, but it basically boils down to extracting the information each indicator holds about age, and omitting from the age estimation other information the age indicators may also measure. With this method, those indicators that are more closely tied to age are given more weight in the final combination of age estimates. Lovejoy and colleagues found that when testing their summary age that it performed with greater accuracy and less bias than any of their component age indicators used individually (1985a). They also found high degrees of underaging in older individuals which they attributed to the individual tendency of the osteological methods to underage as well. As a result of these findings, they introduced a correction step: if the age estimate is over 45 years, they discarded the lowest age indictor and recalculated; over 55 years they discarded the lowest two age

indicators. This step appeared to effectively remove the bias they had observed in the upper ranges of their data (Lovejoy et al. 1985a).

A slightly less statistically-intensive approach to combining multiple age indicators for a single age estimation was suggested by Baccino and colleagues in 1999. Their investigation focused mainly on the preponderance of interobserver error rates (a topic which will be discussed in greater length later in this paper). However, they also tested two multi-trait methods. Working with a sample of 19, they started with the Suchey-Brooks pubic symphysis, the Isçan rib ends, the Lamendin tooth root translucency method, and the Kerley femoral cortical remodeling methods. For their first multi-trait method, they calculated simple averages of the four age estimates. Their second method is slightly more complicated and is similar to the steps of the "complex" method of Acsádi and Nemeskéri (1970). In their method, they used the Suchey-Brooks estimate, but only for the first three phases. If the individual was judged to be in older Suchey-Brooks phases then the age was estimated using the Lamendin method. As mentioned above, these are very simple multi-trait aging approaches, but they did prove to be superior to the four methods individually with respect to accuracy, bias, and observer error (Baccino et al. 1999).

In 2002, Boldsen and colleagues published on a different approach to age estimation. They suggested that anthropologists should stop looking at age indicators and try to define the periods of life during which a skeleton expresses a particular stage. Instead of the old methods, Boldsen et al. looked at defining the age at which skeletons tended to pass from one phase of an age indicator into the next. Their approach, termed "transition analysis," focuses on calculating probability graphs for the age of transitions for all the various transitions different age indicators undergo. Boldsen and colleagues developed their method from a sample of 186 individuals from

the Terry collection split between the four major ancestry-sex groups (i.e., black-white and male-female). In this study they looked at the pubic symphysis, the auricular surface, and cranial sutures. They adopted the approach of McKern and Stewart (1957) and defined components of the indicators: five for the pubic symphysis, and nine for the auricular surface. In addition, they picked five sutures from the cranium to score. The number of different expressions varied between components ranging from three to as many as seven (Boldsen et al. 2002), hence, they did not define any binary aging traits. The stages defined for each trait are assumed to occur in only that specific sequence, so that every individual, should they live long enough, would go through each of the stages in the same order. The only thing that would vary would be the ages at which different individuals would pass on to the next stage of expression. The approach Boldsen and colleagues outlined is rooted heavily in statistical theory pertaining to maximum likelihood estimation and distribution approximation, a topic which will not be explained at this juncture. For detailed equations and mathematical deductions, please refer to their work directly (Boldsen et al. 2002). Using their reference sample, the authors calculated curves of transition that show the probabilities of making the transition from a particular stage to the next for a specific age. These transition curves were mathematically combined into probability density functions, curves that show the age estimation in a probability curve and can be used to derive a point estimate and confidence interval. Boldsen et al. found that the correlations between the actual ages and their estimated ages are high, best for the pubic symphysis, followed by the auricular surface, and trailed by cranial sutures, at 0.86, 0.82, 0.66, respectively. The correlation for all three indicators together was found to be 0.88, leading to the conclusion that the pubic symphysis is the best of the three indicators and basically as informative as all three indicators together (Boldsen et al. 2002).

While the concept of using multiple age indicators to arrive at the best possible age estimations has been around since the inception of age estimation, only in recent history have methods been suggested regarding the ways to objectively combine the information contained in numerous age indicators into a single age estimate. Ranging from simple averaging to complex mathematical procedures only possible with the power of today's technology, they all tend to reinforce the idea that several age indicators are better than one.

Problems in Age Determination

Possibly the greatest difficulty facing the field of age estimation is the age indicator itself and several issues pertaining to it. None of the age indicators that has been studied and utilized to create methods is perfectly correlated with chronological age. From early life the changes to the skeleton are developmental. These changes tend to follow a fairly predictable pattern and the age at which these changes take place is not very variable. Therefore, age prediction based on developmental changes in the skeleton are much more reliable and allow for more accurate age estimates based on analysis of tooth eruption, epiphyseal fusion, and bone growth (Scheuer and Black 2000). However, more developmental changes in the skeleton occur by the time an individual reaches early adulthood. Age-related changes during adulthood, especially in later life, are mostly degenerative in nature and much more variable as to when they take place. As it is true that individuals grow older in appearance at different rates, so it is also true for the skeleton. A multitude of factors can cause early degenerative changes to occur as well as delay them beyond the years at which one might typically see them. It follows then that degenerative changes to the skeleton are much less useful in establishing a tight age estimate, because they have such a high variability. All the age indicators currently in use for age estimation are

degenerative in nature after a certain point has been reached and thus current age estimation grows increasingly less reliable as age increases.

The way the age indicators are defined is also an issue. The two schools of thought on the issue are the phase and the component approaches. We have already seen that both approaches are well represented in the collection of methods currently in use and under development. The phase methods look at the whole indicator and consider it as a single aging variable. Contrarily, the component methods start by looking at the age indicator, identifying individual traits that make up the indicator, and use these traits or components as their aging variable. For each of the components they come up with a set of expressions that are defined with only that particular trait in mind. Once scored, the age indicator's state can be expressed as a string of digits, such as "1134," where each digit represents a specific component and its corresponding stage of expression. To get their age estimate the components get combined, commonly by simple addition, and the component sum is correlated with chronological age. A phase method, on the other hand, provides a summary description for each defined stage of age progression of the age indicator. These summaries, however, take the form of a list of expressions of different aspects of the age indicator. For example, the description for Suchey-Brooks phase 3 reads:

> "Symphyseal face shows lower extremity and *ventral rampart in process of completion*. There can be a continuation of fusing ossific nodules forming the upper extremity along the ventral border. Symphyseal face is smooth or can continue to show distinct ridges. Dorsal plateau is complete. Absence of lipping of symphyseal dorsal margin; no bony ligamentous outgrowths." (Brooks and Suchey 1990)

So, technically, the description reads like a decoded version of the digit code expression of a component method. Yet, if they are looking at the same thing, what makes these methods different? In a phase method, the analyst gets leeway on what he or she looks at to define a

phase. As the phase description includes a large set of traits (not all of which may be available for examination), the analyst will define the phase based on the overall "feel" of the age indicator, weighing in their mind the different expressions of traits. Due to this intuitive aspect of phase methods, they tend to become better with practice and experience. Experienced analysts have a better feeling on how to determine phase of individuals that express characteristics of more than one phase. They know which traits to give more weight when determining phase and may even use additional characteristics not provided by the phase descriptions. This may not be the intent of phase methods, but due to their holistic approach to the age indicator and the inherent variability of humanity, age indicators will express traits from multiple phases. Most methods do not provide the means for dealing with such cases and thus it is up to the analyst to make the distinction. And so the rules set down by the phase definitions become more like actual guidelines.

Component methods are designed to remove the subjectivity for which the phase methods allow. By forcing the analyst to only consider one component at a time this approach seeks to limit the amount of information that needs to be considered to make a determination. Despite this fact, previous component scores may still influence an analyst's determination of subsequent component scores. Yet, the reasons that make component methods less subjective are also a drawback. Since the method defines everything that is examined, no additional information that may be present on the age indicator is recorded or considered. Only the information in the components is considered for the age estimation, which can be considered a good thing or a bad thing.

Interestingly enough, I can illustrate a drawback to each type of method with a single example. In their report, McKern and Stewart used a three-component method each with six

stages on the pubic symphysis. In their sample they found only 21 sets of component combinations and state that probably only about a quarter of all possible combinations are likely to be observed in reality (1957: 83). It seems like they may have forgotten one of their stages. Each of their three component has 6 stages, but since they are numbered 0 to 5 it is conceivable that they may have calculated their total number of possible combinations as 5x5x5. In this case, they would calculate a total number of 125 possible combinations. A quarter of 125 is about 31. When the authors state that they observed 21 combinations and stipulate that about a quarter of all possible combinations are possible, it makes sense if they consider 31 to be that quarter. However, their three component method with six stages each actually has 216 combinations—not 125—of which a quarter would be 54. In this case, 21 observed combinations represents less than half of "likely" combinations, putting in doubt the representativeness of the sample with which McKern and Stewart (1957) were working and subsequently, the representativeness of the method as well.

Barring the erroneous math presented, the component method does have some theoretical strong points. Due to its structure, the component method allows for potentially 216 different pubic symphyseal expressions to be identified and aged. The Todd (1920) phase approach allows for 10 and the Brooks and Suchey (1990) method for only six. Even if it is extremely unlikely that all 216 combinations would ever be recorded or are even biologically possible, the number they did find (21) and the number they postulate as likely to exist (31) are much higher than the number of different pubic symphysis groups defined by any phase method.

Nevertheless, the fact that they only observed such a small percentage of combinations emphasizes that there is correlation between the three components they used. A dorsal score of 5 is likely to be found next to a high ventral score as well. This correlation between components

makes the use of the data for statistical analyses much more complicated and invalidates the assumption of component independence required for some methods.

Mathematically speaking each approach has advantages over the other. Since they are more rigidly defined, component methods produce lower rates of observer error. However, phase methods do not have to deal with correlation between components since they only have one. As all the "components" in "component methods" are defined because they change with age, thus making them useful, it has to be expected that they will correlate with each other. This correlation can cause major problems in statistical analysis; an issue older component methods fail to address.

Anthropologists themselves also exhibit preferences for either phase or component method and this choice can be substantiated for both sides. Component methods attract anthropologists because they allow the assessment of individual characteristics one at a time. The same researchers might shy away from the phase methodology because they find it too limiting or too unspecific. Phase methods can be considered limiting because they only ascribe one score to an age indicator, limiting the amount of information garnered from the indicator. The methods can be characterized as unspecific because they look at the age indicator's general expression and what stage it most closely resembles.

On the other hand, some anthropologists may be attracted to the phase methods precisely because of the freedom these methods allow. They may focus on the great variability in the expression of age related changes and prefer to identify the general state of an indicator, rather than break it down into specific parts that may not agree with one another. In addition, component methods may be viewed as too specific or also limiting. These methods limit the viewpoint that the investigator takes to the skeleton. By focusing on individual traits it is easily

possible for a major general feature to be overlooked. Age estimates based on component methods are very closely tied to the math and how the trait scores are combined and converted to age estimates. Such a rigid age determination may not appeal to some anthropologists who may feel they get a better impression of the age of a skeleton based on a perusal of overall features, rather than individual traits.

The choice between the two approaches presents an interesting dilemma as each has its own advantages and drawbacks. The leeway allowed by phase methods as well as the rigidity of the component methods can be seen as both a blessing and a curse.

Variation in Aging

One of the major overarching problems facing the field of age estimation is the great range of variation expressed by the human species. While other species may be quite varied as well, the amount of study and work that has gone into the study of the human condition has made it blatantly obvious that *Homo sapiens sapiens* is an extremely varied species. The variation is expressed in almost anything, such as stature, skin, hair and eye color, so it should come as no surprise that the skeleton also expresses a large range of variation. This attribute presents a problem to age estimation due to the variation in the expression of age indicator states. It is quite frequent that for the same suite of age indicators, different age estimations are calculated (Boldsen et al. 2002). This variation influences several different regions of the age estimation area of study. These areas can be split into two groups of interpersonal variation (which is discussed below), and intrapersonal variation (which will be covered later in this paper).

Interpersonal Variation

Interpersonal variation pertains to any differences between individuals, be they from different countries and cultures or even just neighbors. Typically, interpersonal variation is studied on the scale of large, distinct groups, as the high degree of human variation makes it too difficult to establish with certainty the characteristics of aging for smaller groups. While there is "interpersonal variation" between a couple of people walking down the street, the sample size would not be large enough to be able to determine the characteristics of the two groups to which these people belong. Therefore, when discussing interpersonal variation in terms of age estimation methods, the groups under comparison must be large enough so that they can be studied as a group and a typical pattern of aging can be established. For age determination purposes, the major sets of interpersonal variation are populations (cultural, native, or ethnic groups) and sexes. Boldsen states that member of different ancestries and sexes are often assigned different age estimates even if the same set of age indicators is used (Boldsen et al. 2002). For example, in the United States, the major groups of interest would be males, females, and members of Asian, European, African, and Native American descent. Further subdivisions can also be defined (i.e. on the basis of economic status) in the study of cross-group differences in the aging progress. All of these groups potentially age differently and could potentially require individual age estimation standards.

Population differences

As of this point no population-specific age estimation standards have been established. Most studies on populations without specific aging standards are treated more like validation studies. Using a new sample, they test the appropriateness of a particular method on a population

for which it was not developed. Rather than develop a new method or a new standard for the new population, the aging pattern of the new population is simply described in terms of the original reference population. In these cases we tend to get very general statements of the amount of inaccuracy and bias that was exhibited by the new population when measured on the standard for the old one. For instance, Hanihara (1952) found that using Todd's ten-phase system assigned ages to a population of Japanese individuals that were on average 2 to 3 years older than the actual ages.

So far, cultural/regional/ancestral differences have not been considered severe enough to warrant creation of population specific methodologies. The extent of a population specific aging methods up to this point simply take the form of "apply method X, and add (or subtract) Y years from the estimate to get your age estimate for the new population." In reality, the "population specific standards" boil down to minor adjustments of existing standards.

The fact that no actual new methods are developed for new populations is due to several intersecting facts. First, the differences between population aging patterns appear to be minor. Second, the patterns of aging appear to be consistent across population groups, while it is the timing of the age-related changes that varies between the populations. So, the same macroscopic evaluation can usually still be applied to the same age indicators of every population. The only things that have to be altered are the age estimation equations and calculations. Most population-specific alterations come in the form of an overall bias adjustment. This assumes that the whole population undergoes the age changes outlined by the reference method at the same speed, just slightly delayed or accelerated. What these simplistic method adjustments fail to take into consideration is that while populations may be going through the same age-related changes, the speed at which they change may not be constant across populations. Some age-related changes

may occur earlier, while others may occur later. (This appears to be the case in many studies, a topic that will be discussed in a subsequent section.) In this case a standard one-number bias adjustment will only average out all the differences into one bias that may actually not create a more accurate age estimation method. Furthermore, it is possible that some populations may lack certain diagnostic age indicator traits. This is true of Isçan et al. (1987) who found that African descendants exhibited much less scalloping on the sternal rib ends, a feature key in age determination. As with all the other problems in age estimation, the problem of sample size is again a major factor in the establishment of new population standards. Most studies and even most collections are just not large enough to allow for the establishment of statistically sound population standard for age estimation.

Pubic Symphysis

Todd (1921) was one of the earliest to look at ancestral differences in the age progression. Using his own ten-phase system he compared his finding on an African American sample to his previous work on people of European ancestry he used as the basis for his system. Todd did observe differences between the two groups, primarily in the expression of some age-related traits and the onset and secession of some phases. In his African American sample, for instance, the symphyseal rim began forming ten years earlier than in the European sample, but unlike the white sample, rarely finished forming completely in his African American sample. However, he did not focus on these differences and concluded that the two groups were essentially the same in their pattern and timing of age progression (1921). Several subsequent works also failed to find significant differences between age determination using Todd's system for groups of African and European ancestry (Meindl et al. 1985; Meindl and Lovejoy 1985).

Kimmerle and colleagues (2008a) looked at populations of American and Balkan peoples and found differences in age changes of the pubic symphysis evaluated using the methods of both Todd (1920) and Brooks and Suchey (1990). The differences between the two populations were not significant, Kimmerle et al. still created Balkan standards of age intervals within the Suchey-Brooks phases (2008a). In a similar study, Godde and Hens (2012) found differences among three populations, discovering that the ages at which Balkan, Sardinian, and American samples progressed through the Suchey-Brooks phases differed between the populations.

Researchers who did find differences between populations found only minor ones. As previously noted, Hanihara (1952) found that using Todd's system established from a white European sample estimated ages for a Japanese sample on average 2 to 3 years older than actual ages. McKern and Stewart (1957) found a greater rate of retarded age progression in their small Black sample, which would lead to underestimation of age for these individuals.

Auricular Surface

The auricular surfaces, similar to the ribs, have seen little investigation into population differences. Using phase stages the auricular surface has yielded no significant differences on the basis of ancestry (Murray and Murray 1991; Osborne et al. 2004). However, Schmitt and colleagues performed an extensive analysis using several continent-defined samples. They found that Asian and African samples do not classified well compared to their European sample. Finding that age-related morphological changes differed between their samples, the authors advocated for the need of population-specific age estimation standards (Schmitt et al. 2002).

Sternal Rib Ends

The pattern found in the study of sternal rib ends mirrors that of other age indicators. Isçan et al. (1987) found only minor differences between samples of people of African and European descent. In the early years and phases no differences were observed on the basis of ancestry. Later phases showed that Blacks tended to develop age-related changes in rib ends earlier than the Whites and were also more varied in the timing and expression of their age-related changes. Overall, the authors concluded that the differences were minor and that it would be tricky to define actual differences between the two populations. The authors suggest that potentially it would be better to develop separate standards for Black individuals (Isçan et al., 1987). In a separate study examining morphological changes of the first rib, Kunos and colleagues also found no influence on age estimation by ancestry (1999).

Cranial Sutures

Todd and Lyon's 1924 study found no differences in suture closures due to ancestry. In subsequent works (1925b; 1925c), they investigated ectocranial and endocranial suture closures separately, determining that only a few endocranial differences in the timing of suture closure existed between the White and Black population samples. These differences were suture-specific and did not establish an overall pattern of population difference even though endocranial suture closure in Blacks showed greater variation than the white sample (Todd and Lyon 1925b). Ectocranial sutures were more erratic and varied in their closure rates, and exhibited slower closure rates than endocranial sutures; however, no overall significant differences were found between the two populations (Todd and Lyon 1925c). Galera and colleagues (1998) substantiated these findings with their own work. Looking at both endo- and ectocranial sutures, the authors

found higher variation in the external sutures with all cranial suture methods showing basically the same results. Contrary to Todd and Lyon's findings however, Galera et al. found that the closure rates in their Black sample more highly correlated with age than their white sample. They concluded that cranial sutures do not lend themselves to accurate age estimation, but they do exhibit population differences in suture closures on the basis of ancestry (1998).

Summarily, population differences in age-related changes have been observed since the beginning. Initially these differences were considered minor and did not need separate methods for each population. Recently these differences have become more noticeable, prompting more anthropologists to consider the need for population-specific age estimation standards. The lack of agreement on this issue illustrates the wide variation that anthropologists encounter in age estimation. Depending on the choice of populations and methods used, some studies will find no differences between populations while others will find differences large enough that the methods may only work well on a single population.

Sex differences

Differences in the aging patterns between males and females have also been observed, although this difference is much easier to study due to the distinct line along which males and females can be separated. Where populations represent a continuous gradient, the sexes represent a dichotomous one, generally much easier to separate for individual investigation. Nevertheless, as with population specific age estimation standards, adequate sample sizes are often difficult to find. From what has been studied on the subject, it appears that males and females by and large appear to undergo the same age-related changes and in the same sequence. The one advantage

that sex comparisons have over population comparisons is that the differences between the sexes are large enough to warrant establishment of separate guidelines.

Differences that are observed between the sexes are attributed to hormone levels (Işçan et al. 1985) or child-bearing (Jackes 2000). Due to the additional factors that child-bearing presents, one might expect age assessment to be less accurate in females based on the greater range of variation associated with the number of parities (Kemkes-Grottenthaler 2002). At least one study (Hoppa 2000) did not find any age change related differences between high- and low-birth females, suggesting that child-bearing has limited (if any) influence on the commonly used age indicators.

Nevertheless, not all studies differentiate between the sexes. If the differences between the sexes are not statistically significant, mathematically speaking there is little reason to establish two nearly identical methodologies. Whether or not a difference is observed between the two sexes appears to be dependent on the age indicator used and the population being studied. For the pubic symphysis, Todd (1920; 1921), Brooks and Suchey (1990), and McKern (Gilbert and McKern 1973; McKern and Stewart 1957) established two separate methods for the two sexes; methods that really do not differ significantly from one another. The number of stages and components are the same and the descriptions for each are synonymous. As these methods differ in the estimated age ranges associated with each of the phases/component summary scores, these age estimation standards for females are basically phase-specific bias adjustments.

Pubic Symphysis

Todd followed his groundbreaking work on the morphological changes with age on the male pubic symphysis by looking at the female symphysis as well (1921). He found that females undergo age changes in the symphysis much like males do and defined ten phases for the female age progression, just as he did for the male progression. The main differences Todd noted pertained to the morphological structure of the symphysis which were not necessarily related to age changes. He noted that in females, the pubic tubercle is set back from the symphyseal surface, and the surface had more irregular outlines and were more sharply lipped than male symphyses. The actual progression through the ten age phases showed only slight differences between the sexes. Females tended to be delayed about two years in phase three, but were equal to males by the end of phase seven. Todd (1921) finished his study by concluding that there were more differences between males and females than between his samples of Black and White individuals.

In the follow up study by Gilbert and McKern (1973) in which the authors attempt to address weaknesses they observed in Todd's method, they also noted sexual differences when utilizing an age component system. The authors point out multiple differences between their sample and that of the comparable male study (McKern and Stewart 1957), potentially inflating significances in order to increase the validity of their female aging method. Like Todd, they found structural as well as age-related differences, however, they were not the same as those observed by Todd. Structurally, they suggested that the ventral and dorsal aspect of the pubic symphysis in males is separated by an imaginary line while in females the separation is delineated by a beveling of the ventral rampart away from the dorsal surface. Following this difference the authors further observed that the symphyseal rim surrounds the whole surface in

the males while it separates the two demifaces in females, circling the dorsal rampart. In terms of timing of age changes, Gilbert and McKern found that the ventral rampart, rim, and margin develop about four years later in females than in males. Conversely, the dorsal surface flattens faster in females by about 7 to 10 years (Gilbert and McKern 1973). Finding primarily morphological differences in their own study, Meindl and colleagues (1985) concluded that differences due to sex were not significant. They observed that female symphyses tended to be smaller and the smaller pubic symphyses yielded less accurate age estimates than large symphyses. Further, the lower extremity of the symphysis is easier to identify in females than in males due to the associated morphological shape of the inferior pubic ramus (Meindl et al. 1985). Katz and Suchey (1986) found than females show greater variability in expression than males. Later, Brooks and Suchey (1990) also found higher standard deviations for females which they attributed either to higher variability or greater age range for the female sample. They also suggested that the female pubic symphysis ages primarily in the ventral aspect, dorsal changes are not tied to age, and dorsal lipping is not a female age trait.

Using the McKern methods (McKern and Stewart 1957; Gilbert and McKern 1973) as part of a validation study, Jackes (1985) found that these methods yielded more accurate results for young males and for older females. Saunders et al. (1992) found that the Suchey-Brooks method produced better results for females. Hens and colleagues (2008) stated the same thing, finding estimates for females based on the Suchey-Brooks method to have lower overall bias than males. They also found that females showed no bias until age 29, while males fared better at not showing bias until ten years later at age 39. Females are more subject to osteopenia and osteoporosis and thus would age faster and appear older than males until age 80, when the two sexes are the same in terms of pubic symphysis morphology (Berg 2008; Kimmerle et al. 2008a).

Kimmerle and colleagues (2008a) found significant differences between samples of American and Balkan women when using the Suchey-Brooks method. Since the populations as a whole are not different, the differences observed among the females are attributed to health, diet, or biological factors such as osteoporosis.

Auricular Surface

Looking at the characteristics used to define age-related changes in the auricular surface, Lovejoy and colleagues (1985b) found no sexual differences to their characteristics, including apex, striae, billowing, macroporosity, transverse organization, retroauricular surface, grain, and density. They did find that females having a preauricular sulcus show increased changes in the apex and inferior margin of the auricular surface and suggested that these two sites should be excluded from age evaluation for females exhibiting preauricular sulci (Lovejoy et al. 1985b). Murray and Murray (1991) found no differences on the basis of sex after adjusting the method for conflicting results. Other researchers (e.g., Buckberry and Chamberlain 2002; Osborne et al. 2004; Schmitt et al. 2002) also found no sexual differences in their auricular component scores or age determined from auricular surface morphology

By contrast, Igarashi et al. (2005) found an interesting difference between the sexes. Modeling auricular surfaces from component scores, they found that the male equations had the best results overall. More interestingly, they found that the equation for pooled sexes worked better for young females than their female only equation, while the female only equation worked better for older women than the pooled equation. They attributed this peculiarity to the sexes having different modes of chronological change. Hens and colleagues (2008) found that the Lovejoy phase method produced greater bias and inaccuracy for males than females.

Sternal Rib Ends

Looking at the morphological changes in sternal rib ends for females, Işçan and colleagues (1985) outlined the differences between his female and male samples in detail based on the standards established from the male sample. Compared to the males, female ribs began changing earlier, with further development continuing to stay ahead of comparative male development until 28 years. After that age, females slowed down their morphological aging and were one to five years older than males for the same phases. Purely morphologically, Işçan and colleagues observed female ribs were thinner, less dense, and grew smaller body projections with shallower pits than males. They also formed pits earlier and grew bony projections much later than males (1985). A study of the age-related changes to the first ribs did not find significant differences between the sexes (Kunos et al. 1999).

Cranial Sutures

Much like other studies on the range of human variation pertaining to age estimation indicators, studies on cranial sutures have returned various results. Multiple sources (Acsadi and Nemeskeri 1970; Meindl and Lovejoy 1985; Todd and Lyon 1924) claim there is either no difference or could find no difference between the sexes in their studies of the rates of cranial suture closure. Ashley-Montagu (1938), on the other hand, cited evidence to support that males fused on average earlier than females, and Meindl and colleagues (1983) similarly observed that suture closures were highly variable and differed between the sexes. Brooks (1955) stated similar findings but added a range to the observed differences. According to this study, female cranial sutures closed 5 to 25 years later than male cranial sutures. In a skeletal sample from Spitalfields, Key and colleagues (1994) found male cranial sutures closed faster than females. They also

observed that differences between the sexes were more prominent for ectocranial sutures while endocranial sutures showed little difference in closure rates between males and females. Galera and colleagues' (1998) study found no sexual differences in suture closures and that the Lovejoy-Meindl method proved more accurate for males than females.

Intrapersonal Variation

Intrapersonal variation refers to the variation that exists within a single individual, such as between multiple age indicators. Intrapersonal variation could be studied in two different ways. First, age estimates from different age indicators in the same individual can be compared to each other. For instance, the age estimate based on the pubic symphysis can be compared to that of the auricular surface to determine if the two sites agree or disagree and if the pattern of the comparison is consistent across multiple individuals. Secondly, age indicators from right and left sides of the body can be compared to determine if they show the same age-related changes to each other.

Unfortunately, intrapersonal comparisons of age indicators are not feasible. The biggest problem is that the age estimation methods are not good enough. It is not possible to compare a 25-45 year age range based on pubic symphysis morphology to a 40-44 year age range based on fourth rib end morphology. Any age estimation method with a large age estimation interval provides little useful information for within-body comparisons. Therefore, without better age estimation methods, any comparisons between age indicators remain very tenuous.

Todd (1920) does provide a significant number of special cases in his study of pubic morphology of individuals who show abnormal patterns in their skeletons when compared to clinical records. Some show total body retardation or acceleration, where all bones in the

skeleton appear older or younger than they should at that person's age. More interesting are those individuals who show only parts of the body that are older, while others appear as expected. Most of these cases can be explained by medical conditions or chronic living conditions. In either case, they show that a certain level of intrapersonal variation can exist. In other words it is possible for age indicators to disagree with each other or with the actual age of an individual. Hence, without knowing the recent life history of a skeleton it is considered unwise to estimate age on the basis of a single age indicator. Almost every anthropologist who has mentioned multiple age indicators argues for the use of multiple indicators to arrive at more accurate age estimations. While the age estimates may not actually be more accurate, basing an age estimate on multiple age indicators is likely to ferret out any indicator that was unduly affected during life and is providing the scientists with an inaccurate accelerated or retarded age estimate.

While the study of variation within the body is not very feasible for the comparison of multiple age indicators, it is still possible to study intrapersonal variation. By looking at right and left sides of the same age indicator in a single individual, it is possible to study the variation between the two sides. This variation should be less than that between different age indicators, so if some can be shown to exist, it would further justify the argument for the use of multiple age indicators to estimate age. However, the study of age indicator symmetry is of limited interest to the anthropological community. Very few scientists consider it and if they do they only mention it in passing. Worse, most do not even appear to consider it and gloss over the fact that the two sides might not agree with one another. The common practice is to assume that an age indicator is symmetric in its expression and that by using one side consistently any asymmetry is eliminated. However, a common convention is to "use the left side unless it is absent or damaged

and then use the right side." Such studies rarely, if ever, consider the ramifications of potential asymmetry. It is correct that a few asymmetric individuals will be overshadowed by the assumed majority of symmetric individuals, and that any conclusions and results of a study are primarily shaped by data from symmetric individuals. This does assume, potentially erroneously, that asymmetry is rare and abnormal.

As mentioned above, few individuals have considered symmetry, and of those, most investigations were superficial and byproducts of other studies. The following paragraphs outline some of these findings. Plenty of studies have found no significant difference between the left and right sides in binary and ordinal auricular components (Buckberry and Chamberlain 2002; Falys et al. 2006; Igarashi et al. 2005), cranial suture closures (Galera et al. 1998; Jackes 2000; Todd and Lyon 1924; Todd and Lyon 1925a), and Suchey-Brooks symphysis phases (Hens et al. 2008). McKern and Stewart (1957) stated that they found no significant differences between the sides throughout the whole body. While they spotted some asymmetry they did not deem these instances significant enough to elaborate. Schmitt (2004) found a fair amount of asymmetry in about a quarter of pubic symphyses and a third of auricular surfaces. In their research with auricular surfaces, Hens and colleagues (2008) found no asymmetry between male innominates, but did find a significant difference between female auricular surfaces. Moore-Jansen and Jantz (1986) report finding asymmetry in auricular surfaces.

The greatest amount of asymmetry has been reported in the closure rates of cranial sutures. As cranial sutures exhibit some of the greatest variation in suture closure, it is easily possible that patterns observed can be products of sample sizes or composition. Kemkes-Grottenthaler (1996) found that the left side of the cranium closed its sutures more slowly than the right side. Meindl and Lovejoy (1985) recorded asymmetry to be common for the zygomatic

and malar sutures, while Key and colleagues (1994) observed considerable nonsignificant amounts of asymmetry in suture closure.

While surprisingly little work has been done investigating asymmetry as it pertains to age estimation, a related topic has received more attention. The methodology Isçan established for age determination based on the morphology of the fourth rib (Isçan et al.1984a) is quite limiting considering that there are 24 ribs in the human rib cage. As such, studies have investigated if ribs other than the fourth are also useful for age determination using the same standards. No differences were found between the third, fourth, and fifth ribs, thus allowing for use of either for age determination (Dudar et al. 1993; İşcan and Loth 1988). McKern and Stewart (1957) stated that the uppermost and lowermost rib head mature faster in terms of epiphyseal fusion. They grouped ribs four through nine into the same category for head fusion. While not pertaining directly to sternal rib ends, it is possible that the pattern observed at the rib heads can also be evident in the sternal ends. This suggests that the first and last ribs should be carefully considered before being used to estimate age by standards established for the fourth rib.

Observer Error

Observer error is one of the most frustrating factors in age estimation from skeletal remains. We have already outlined the common age determination methods. These methods all use descriptions in the form of paragraphs or sentences to describe the extent of observable skeletal morphological changes. Due to the fact that these classification categories are descriptions, they add a great degree of observer subjectivity to the age indicator evaluation and staging process. In other words, what one observer may classify as “minor billowing,” another observer looking at the same skeletal element might call “absence of billows.” When two sets of

eyes look at the same thing but "see" something different, this is observer error. It is accepted that observations are going to be subjective and therefore have observer error. It follows that reduction of subjectivity will decrease observer error. Hence, long bone lengths which are measured using standard equipment have minimal subjectivity and should have much reduced levels of observer error (Gowland and Chamberlain 2002). The reduction of observer error is one of the major attractions of component methods. The theory holds that component definitions are more restricted and more objectively defined allowing for less subjective interpretation of the age indicator descriptions. If this stipulation holds true then component methods should have a lower incidence rate of observer error than phase methods (Kemkes-Grottenthaler 2002).

Two kinds of observer errors are interobserver and intraobserver error. Akin to interpersonal and intrapersonal variation, the definitions are similar. Interobserver error is what the example above illustrated: two different observers looking at the same skeletal elements and recording something different. In the case of age estimation this would be two age indicator stages or phases that are not the same. Intraobserver error occurs when the same observer looks at the same element multiple times and ends up recording different results. These subsequent observations can happen close together or even several months apart depending on the situation.

The problem that observer error presents is quite obvious. Looking at the same feature should result in the same conclusion being reached. If the observers do not agree, then it becomes uncertain which one of them made the correct or more accurate evaluation. As these observations are generally tied directly to the age estimation, two different observations translate into two different age estimations. Generally, the observations are not diametrically opposed but rather close to each other. Nevertheless, any level of observer error adds additional doubt to a

process already fraught with uncertainty. Given its nature, observer error is a persistent and difficult problem to address in age estimation.

The fact that observer error is a problem in the first place is troubling. One of the fundamental characteristics of science is its repeatability whereby results can be recreated and verified, yet observer error is direct evidence that age estimation is not repeatable. Sure the process is repeatable, but the results will not always be identical. Probably the greatest contribution to the amount of observer error is the high degree of subjectivity, mentioned earlier, that comes from using descriptive age indicator categories. Subjectivity breeds observer error. Several aspects can cause error. Unclear wording is likely to lead to varied interpretation of the text. Even succinct text can be interpreted differently by anthropologists when looking at skeletal specimens. Furthermore, the longer the description of a phase and the more aspects a description addresses, the more likely that scientists will focus on different parts of the description and potentially end up evaluating skeletal age indicators differently (Schmitt 2004). For example, the Suchey-Brooks method describes six phases of the pubic symphysis. Each description is a small paragraph and addresses numerous features observable on the symphysis including dorsal bevel, ventral rampart, billowing, rim formation, erosion, extremity formation and so on. When trying to identify the phase to which a particular specimen should be assigned, anthropologists may go about it differently. They may rely on features that they can more easily identify than others, or may only look at features that they are comfortable identifying in the first place, disregarding features that are unclear or they find ambiguous. One might expect that the bigger the description the more subjective it is, and the more prone to observer error it becomes.

Subjectivity also factors into a second contributor of observer error: the experience of the observer. Every observer will bring a level of experience bias with them. The populations on

which an anthropologist typically performs age estimations will determine their approach to new specimens. Someone who has only seen very few young skeletons will be more uncertain about aging a young skeleton than older ones with which they may have extensive experience. Aging one individual after another also introduces error due to the changing standards over time (Lovejoy et al. 1985b). Experience level and subjectivity certainly go hand in hand to a certain respect and both are major contributors to the presence of observer error.

While troubling when present, observer error is not difficult to deal with. Ideally, for interobserver error the observers should discuss their results and justifications for assigning different values and come to a joint conclusion. For both types of observer error, a second (or potentially third) look at the specimen is warranted. If no agreement can be obtained, the only resolution is to create multiple age estimations and hope they are close enough that they both fall in a common age estimation interval, which is the common presentation of age estimation results.

Observer error is a common occurrence and as such it is the severity that is generally of interest. Many studies may dismiss small amounts of error and presentation of results is difficult as both the frequency and the magnitude of the error is of interest. If either the magnitude or the frequency is low enough the "amount" of observer error can be dismissed as constituting a nonsignificant amount.

Interobserver Error

Results for studies investigating interobserver error vary quite widely. Some studies find no or only very small rates of interobserver error. Such results have been found for analyses of pubic symphyses (Berg 2008; Hanihara and Suzuki 1978; McKern and Stewart 1957), auricular

surfaces (Buckberry and Chamberlain 2002; Lovejoy et al. 1985b), and histological methods (Baccino et al. 1999). Other studies have found significant levels of interobserver error looking at the zygomatic and malar sutures (Meindl and Lovejoy 1985), pubic symphysis components (Suchey 1979), and pubic symphysis phases (Saunders et al. 1992).

Baccino et al. (1999) found significant differences in the assessments of ribs and pubic symphyses for two observers. The error was evident both in the direction and size of the age estimation biases. Kunos and colleagues (1999) found that while interobserver differences were not significant overall, on an individual basis the differences could be quite large. Using the summary age method of estimation (Lovejoy et al. 1985a), Bedford and colleagues (1993) observed low levels of interobserver errors, with the largest coming from evaluation of the pubic symphyses. On average these errors came to approximately 5 years between observers. Isçan and Loth (1986) briefly mention interobserver error in their validation study. While their results showed low levels of accuracy on the part of the various observers, the interobserver differences were actually quite low. In that particular case, the observers were closer to each other than to the true age of the specimen, indicating a problem with the method and not interobserver error.

Galera and colleagues (1995) looked at various age indicators and found no significant interobserver error. Sutures showed a small amount, with ectocranial sutures exhibiting more than endocranial sutures. They also found high levels of interobserver error for evaluations of the ribs and pubes, but mathematically these levels were not classified significant. Their results confirmed that methods with fewer categories exhibited lower levels of disagreement between observers.

Kimmerle et al. (2008b) also found large rates of disagreement between four observers. Using Suchey-Brooks (1990), Todd (1920), and Isçan (1984a) phase methods, they found that

agreement between the observers was rare. The rib method performed the worst, followed by the Todd and then the Suchey-Brooks method. Further evaluation showed when it came to looking at teeth, the observers were in total disagreement. This problem was not alleviated by calibration as the observers still did not reach consistency when working independently. Overall, the authors found that most interobserver errors were a single phase apart and that observer error more commonly occurred in the middle phases.

Intraobserver Error

In several studies (Buckberry and Chamberlain 2002; Falys et al. 2006; Murray and Murray 1991; Nagaoka and Hirata 2008), the auricular surface has yielded no evidence of intraobserver error for either the Lovejoy phase or the Buckberry-Chamberlain component methods. Still other studies have also found little to no intraobserver error for pubic symphysis phase seven (Berg 2008) or first rib morphology (Kunos et al. 1999).

Dudar and colleagues (1993) found high levels (34%) of intraobserver error using Isçan's phase method. Likewise, Saunders and colleagues (Saunders et al. 1992) found moderate to very high levels of intraobserver disagreement for pubic symphyses (19.3%), sternal rib ends (28%), and ectocranial sutures (40-60.7%), where the sutures of the vault were more consistently evaluated.

Method Validation and Reliability Testing

Possibly the most common complaint about aging methods is that they do not perform well when someone uses it themselves. In this section, we will discuss the problems that have

been revealed by testing of methods; problems both with the methods as well as the age estimation process itself.

In age estimation, the performance of a particular method has traditionally been measured by the average difference between the actual and estimated age. To this measure of accuracy, analysts have added the measure of bias, which measures not just the inaccuracy of the estimates but also takes into account the directionality of the inaccuracy. Where accuracy measures the tendency to miss-age an individual, bias measures the tendency to over- or underage an individual. These two measures can be calculated for individual specimen or be averaged over a whole sample. A third measure of performance, correlation, only applies to samples and measures the strength of the relationship between two variables. In our case, correlations would most commonly be calculated between estimated age and actual age, or between an age indicator and actual age. Correlation values range from -1 to 1, and high values, whether positive or negative, are good, indicating a strong association between the two variables. However, due to the way most age indicator stages are defined, a positive correlation with age is expected. Since age stages are defined in the chronological order that they are presumed to occur, these age indicators have a built in positive correlation with age.

An aging method is developed from a reference sample and all the parameters, be they age estimation equations, phase means and age ranges, or probability curves and density functions, are calculated from this reference sample. A reference sample is chosen to represent a larger population. As such, it is presumed that variables estimated from the reference sample will also work for any sample drawn out of the population. This can be tested using method validation, the process of using an established method to testing a sample different than the reference sample from which it came. This procedure is taken from the field of statistics where a

validation sample is expected to perform worse than the reference sample precisely due to the way statistics are calculated. While there may be some statistics that could turn out better, by and large the validation sample, taken from the same population, will not surpass the performance of the reference sample when both are tested using the method established from the reference sample.

The performance of a method under validation is thought to more closely approximate the true performance capabilities of the method. For example, Lovejoy and colleagues (1985a) tested their method on their sample and got an average inaccuracy of 7.3-7.8 years and a bias of -0.5, but then Osborne and colleagues used the Lovejoy method on a sample from the Terry and Bass collections and found an average inaccuracy of 11.4 and a bias of 1.2 (2004). In this case, the age estimation method defined by Lovejoy should only be credited with the level of performance seen in Osborne's validation sample. But what does this mean? We know that the validation sample will not perform as well as the original, so a small difference is expected. If the validation sample performs very poorly or if there is a large disparity between two samples, then something is amiss.

Sample Appropriateness

The difference between the two samples can be caused by a variety of problems. If the age indicator is only loosely tied to age, a second sample may have a completely different age-relationship to the same indicator. Inexperience on the part of the researcher can also lead to misapplication of a method, resulting in different performance of the method. Misapplication can also result from poorly defined age indicator stages, badly worded descriptions, or any other factor that increase interobserver inconsistency.

The most common explanation given in the literature to explain poor performance of an age estimation method in a validation study is the inappropriateness of the validation sample itself. Most age determination methods are developed based on a readily available sample that is supposed to represent a large population, say all of humanity. Obviously a universal method of age determination would be more useful than one that only works for Europeans or Native Americans and is therefore the kind of method anthropologists strive for. Since the reference sample represents only a fraction of the population, it is hard for a reference sample to cover the complete range of variation expressed in the population. Hence, characteristics based on the reference sample will be more specific to the reference sample than the population as a whole. If these characteristics are age-dependent, any aging method may become too sample-specific and may skew age estimations for individuals from the population that fall outside or near the edges of the range of variation of the reference sample. This is the primary reason for poor performance of method validation: the validation sample comes from a categorically different population than the reference sample and therefore is not an appropriate sample to test the given method. The originators of a method rarely constrain the applicability of their method, because they want and hope that their method will in fact work for all populations. However, once someone comes along and proves that the method does not perform well on some other sample, from then on the validity of the method is questionable in its application to any population samples other than those from its reference population.

Nothing can really be done to address this issue from the side of the aging method. Once a method is defined the only thing left to do is to test a variety of different samples and define the broader population for which the method can be applied with reasonable success. To this end an acceptable level of disparity between validation and reference sample needs to be defined and the

different validation samples should test both the regional as well as the temporal range of applicability of the method. If an aging method is only applicable to a very small population, it is obviously of very limited use. From the methodological side two ways could reduce the problems revealed by validation studies. First, the reference sample should be large and cover a broad range of individuals, so as to be as population non-specific as possible. In this case, any validation disparities are likely artifacts of the validation sample. Alternatively, the parameters of the age determination method could be defined more broadly. However, due to the nature of the aging process, any such method may end up being too inaccurate when it comes to the age estimate, and thus be quickly forgotten by the anthropological community. Of course as a researcher, it is always possible to just stick with what you know: the reference sample works and any validation sample that says otherwise indicates a problem not with the method, but rather with the process with which the validation sample itself was analyzed.

Nevertheless, even without the definition of an appropriate population, poorly validated methods continue to be utilized, most likely because there is nothing better available. What researchers should always keep in mind is that any method is defined by its reference sample, no matter how broad it might be and therefore using a method on your own sample needs to be done with a grain of salt.

Validation tests on the aging methods outlined previously have given varied results. These range from suggesting an alternative method that performs better, to validating a method by showing slightly worse performance on a new sample, to severely questioning the viability of a method. There is more to the process of validation than simply trying out a method on a new sample and calculating the accuracy, bias, and age-indicator correlation for the new validation

sample. If one takes the time to look more closely at the results in the validation sample, certain patterns may become apparent.

Inaccurate Aging of the Elderly

Possibly the most common pattern observed in validation studies is the inability of methods to age old individuals accurately and the progressive worsening of age estimation as age increases (Boldsen et al. 2002; Hoppa and Vaupel 2002). This leads to the inability to assign narrow age ranges to the skeletons of the elderly.

Several reasons for this phenomenon have been put forth. The main reason is that while age-related changes in the young are based on developmental changes, the age indicators used to age older individuals are based on degenerative changes (Boldsen et al. 2002). These degenerative changes are much more easily influenced by a variety of environmental and life factors, thus introducing a higher degree of variation for the expression of the degeneration of an age indicator. Further, the older an individual becomes, the more time is available for the degenerative process to continue and for other factors to influence it, either accelerating or slowing down the degeneration. Hence, if looking at a degenerative age indicator, older ages automatically have a greater range of expression of the age indicator.

Secondly, the accuracy for aging old individuals is low because most age estimation methods were developed on younger samples and are developed with static age estimates. This means that most age estimation methods used reference samples ranging in age from around 16 to 50 or 60. Most of the analysis was focused on the early developmental age changes while little attention has been paid to the degenerative changes age indicators undergo later in life. As one of the main reasons for age estimation outside the field of forensics is the study of early and

ancient populations, their mortality profiles and demographies, age estimation above a certain age is not particularly interesting. For one, early populations generally had much shorter life expectancies and rarely, if ever, reached what we now consider old age. For another, in the realm of fertility, females over the age of around 45 do not need to be aged more accurately as a more exact age estimation would not provide any additional information about the fertility of the individual. Furthermore, many of the samples used to establish the aging criteria are also from populations with age ranges not representative of the whole human race (Brooks and Suchey 1990; McKern and Stewart 1957). The age range of the sample limits the range for which ages can be estimated, leaving any individual outside the appropriate age range likely to be greatly miss-aged. This deflates the average accuracy of the age estimates.

Mathematically speaking, estimating the ages of individuals that fall inside the age range studied in the reference sample will be more accurate. Within the age range of the reference sample, more age-related information is available to help in the age estimation process. Outside the known age range, any estimate would have to be made by extrapolation based on the knowledge within the age range. Generally this process is completing by continuing the pattern seen within the age range and extending it beyond. Often times the pattern is simply considered linear, but more complex patterns are easily possible. Anthropologists are well aware that aging does not take place at a continuously smooth pace, but that there are periods of faster and slower aging. Also, the farther the extrapolation strays from the known data set, the greater the mathematical uncertainty in the prediction becomes. Without knowing the pattern of aging it is not reliable to apply extrapolation to the realm of age estimation.

In addition (and more importantly), extrapolation is pointless to do unless you are working with continuous or ordinal variables. While anthropologists treat age indicator states as

ordinal variables, they may not be. Age indicator states are definitely defined as categories with an implied order. This makes the use of extrapolation very questionable. Therefore, using the current age indicator variables, it is mathematically unsound to estimate anything outside the range of variation observed within the age range of the reference sample. Sample age ranges, variation in degenerative age changes, and inability to extrapolate outside the known frame of reference all contribute to the inability of age determination methods to apply accurately to older individuals and support the finding of greater inaccuracies with increasing age.

Three basic solutions present themselves to address the problem of aging old skeletons. First and foremost, any age determination method has to include old skeletons in their reference sample. If the reference sample age range does not include a fair number of elderly specimens, the likelihood that the method will age elderly accurately is very low. Instead they will be severely underaged. Secondly, age indicators have to be defined for the whole age range and should not stop at some arbitrary age with only "no further changes" or "continuing slow degradation" described after that age. Such definitions may present a problem with the currently used set of age indicators and hence as a third solution, new age indicators that may show more variability in the mature skeleton should be investigated and added to the anthropological aging repertoire. A whole body approach to aging with several indicators useful for certain age ranges may allow anthropologists to become more accurate in the age determination for elderly skeletons.

Regression to the Mean

A more interesting pattern observed in validation studies is a systematic overaging of the young in the sample and consistent underaging of the old individuals in the validation sample

(i.e., Bedford et al., 1993; Jackes 1985; Meindl et al. 1985; Saunders et al. 1992). This phenomenon has been given different names such as “attraction to the middle,” “attraction to the mean,” and “regression to the mean” (Masset 1989). This feature can be a result of a narrow age range of the sample or the methodology used. If an age indicator is only defined from a reference sample with a restricted age range, for example 25-50 years, then any individuals who fall outside that age range cannot be accurately analyzed and will skew the results. Individuals younger than 25 will commonly be grouped together with the youngest defined characteristics and be lumped in with the youngest age group of 25 year olds. Those individuals in the validation sample outside the upper range of the reference sample will likewise be assigned ages like the old specimen from the reference sample. However, due to the restricted age range of the reference sample, a method built upon that is unable to estimate ages lower than 25 or older than 50. Everyone outside the age range will be aged closer to the mean and thus create the pattern of "regression to the mean." An example of a rather restricted age range used in age determination can be seen in the work of Hanihara and Suzuki (1978), which used a reference sample with an age range between 18 and 38 years.

Alternatively Aykroyd and colleagues (1997) have shown that the same characteristic pattern is a direct result of using linear regression to regress ages on age indicator states. Due to the way least squared regression works to calculate a line to minimize error distances to the data points, Aykroyd and colleagues showed that the line through the residuals graphed over age has to be positive. Therefore, young individuals will always appear older than they are and older individuals will always on the average appear younger than they are. Furthermore, they concluded that only a perfect age indicator would avoid this pattern. Since most age indicators

are far from perfectly correlated with age, the attraction to the middle will be present to some extent or another for all imperfect age indicators (Aykroyd et al. 1997).

The simplest solution to avoid regression to the mean is to stop using the inverse calibration method of regression where the age indicator is treated as the independent variable to predict the age. Instead, one should use classical calibration where the age as the independent variable is used to predict the age indicator as the dependent variable based on the reference sample. By inverting the solution found, it is still possible to calculate an estimate of age from the age indicator for future unknown individuals. This method eliminates the "regression to the mean," although the age estimates have a broader prediction interval (Aykroyd et al. 1997). Unless the age-at-death distributions of the target and reference samples are similar, the use of inverse calibration will lead to age estimates that are biased towards the distribution of the reference sample (Prince and Konigsberg 2008).

Age mimicry

Another major problematic phenomenon for age estimation that can emerge from method validation is what is called "age mimicry" (Boldsen et al. 2002; Hoppa and Vaupel 2002; Prince and Konigsberg 2008). This term is used when, after applying an age estimation method to a validation sample, the age distribution of the sample resembles that of the reference sample used to establish the aging method. This is a major problem for anyone interested in estimating the age of unknown individual, be it a single person or a whole population. In the case of demography the problem is readily apparent. An age estimation method that creates a distribution that is not representative of the population under scrutiny—but instead resembles the distribution of the population that was used to create the aging method—cannot be used to infer

anything about the population of interest. That is, unless you make the potentially shaky assumption that your population and the reference population have basically the same age distributions. In the case where age is estimated for a single individual, there is no distribution and so any possible age mimicry cannot be shown; however, with age mimicry, many ages are skewed away from their true ages toward more reference sample-like ages. Age mimicry obviously introduces a certain level of bias in the age estimates which is not something fondly welcomed with open arms when trying to estimate skeletal ages.

Age mimicry has been shown to be the result of improper application of age estimation methods. Holman and colleagues (2002) stated that estimating ages from indicator aging tables and then combining these estimates into an age distribution results in age mimicry. Similarly, Müller et al. (2002) described the origin of age mimicry slightly more completely. Most aging methods correlate a particular age indicator state with a corresponding age range and mean age. Estimating ages directly from these mean category ages based on the reference sample yields very disparate age distributions as such a population only has individuals with x different ages where x is the number of age indicator states. In other words, if you age a person based on the mean age of an age indicator state, every individual with a stage 3 pubic symphysis would be given the same point age estimate (i.e. every stage 3 symphysis belongs to a 24 year old). Using such point estimates to create a population age distribution yields one that is biased towards the reference sample.

Konigsberg and Frankenberg (1992) take a more mathematical approach to the explanation of age mimicry. In their work, they show that using the traditional age estimation approach derived from Bayes' theorem leads to biased estimates in cases of individual age estimation as well as in the estimation of the population age distribution. Estimations based on

Bayes' Theorem work on the marginal probabilities established in the reference sample and translate them to the target population or individual. Bayes' Theorem allows the calculation of the probability that an individual is a particular age given its indicator state. Through their math, Konigsberg and Frankenberg show that the age distribution of the target sample is almost always influenced by the age distribution of the reference sample. The amount of influence is dependent on the strength of the correlation between the age indicator and age itself. The authors concur with Bocquet-Appel and Masset (1982) in that the less related the indicator and age are, the more closely the target sample age distribution will resemble the age distribution of the reference sample (Konigsberg and Frankenberg 1992). Jackes (2000) agrees, stating the inverse, that age estimates will only be accurate when the age distributions of the target and reference sample are similar.

If the Bayesian method is used for the estimation of the target age distribution, then the target and reference distributions will only be independent (i.e. lack any level of age mimicry) if the age indicator is perfect. Again, this is highly unlikely. Konigsberg and Frankenberg (1992) continue to state that by using the Bayesian method the target age distribution will always be biased unless the target and reference sample age distributions are the same, the distribution of the reference sample is uniform, or if the age indicator is perfectly correlated with age. Since none of these three cases is likely, the authors argue for the use of maximum likelihood estimation of the target age distribution to avoid any chance of age mimicry. Maximum likelihood estimation ignores the data from the reference sample and iteratively searches for a distribution that best fits the data in the target sample. Therefore, while the Bayesian method is subject of a certain degree of age mimicry, the iterative maximum likelihood estimation approximates the target age distribution quite well (Konigsberg and Frankenberg, 1992).

Age mimicry can be addressed by the correct statistical methodologies or appropriate choices for reference samples (Prince and Konigsberg, 2008). In cases where an appropriate reference sample is available, the traditional inverse calibration regression with the Bayesian probability calculations can be used. This is commonly the situation in forensic anthropology where the population of the unknown is highly suspect and thus an appropriate "modern" reference sample is usually available. However, in cases where there is no certain knowledge about the age distribution of the population from which the target sample came (i.e. no informative priors), then classical calibration should be used to produce maximum likelihood estimations. In their example, Prince and Konigsberg (2008) support their claims using classical calibration. For their sample, they reduce the overall mean error, aging bias, age mimicry, and age estimation ranges and produced higher correlations between estimated and actual ages than comparable inverse calibration methods utilized by others (Lamendin et al. 1992; Prince and Ubelaker 2002).

Examples of Method Validation

Studies pertaining to age estimation are primarily based, at least to some extent, on previous works. They may adopt parts of the methodology, the idea, the concept, or the age indicator definitions. They may incorporate these portions into their own works, modify them, or expand upon them. In the broadest sense of the word, almost all age estimation studies will be "validating" a previous method to some degree. When Suchey and Brooks (1990) developed their six-phase system based on the 10 phases defined by Todd in 1920, they validated his descriptions of the morphological changes with age of the pubic symphysis. When Buckberry and Chamberlain (2002) defined components for age determination of the auricular surface, they

"validated" the original component methodology put forth by McKern and Steward in 1957. In this way, most studies of age estimation can be viewed as "validating" some other method or concept. Actual validation of a method requires a more stringent approach. Statistical validation of a method is the use of the exact same method as established in the literature on a different sample. Statistically speaking, if the two methodologies (those of the original study and the validation study) differ in any respect other than the data sample, then the results of the study are not statistically validating the original method.

The following examples of validation studies and their results represent only a fraction of the information in the literature. Of note is that validation studies are often the first step in presenting an alternative age estimation method. In these cases, the focus of the study lies in the author's own suggested methodology. Therefore, the initial validation results that prompt the development of the new method are often presented in less than thorough fashion. As an example, Lovejoy and colleagues (1985a) compared their results using several different methods of age estimation on the same sample from the Hamann-Todd collection at the Cleveland Museum of Natural History. They looked at pubic symphyses, auricular surface morphology, femur radiographs, dental wear, and cranial sutures to estimate ages. The authors presented their findings for each of the methods in terms of bias and inaccuracy. Rather than discuss these findings in lieu of the validity of the associated methods, the authors used the results to argue for a new multi-trait method of their own development that outperforms the methods they tested (Lovejoy et al. 1985a). This kind of study is only a partial validation study. It uses established methods to create data, but rather than use the data to discuss and validate the methods, the validation results are used to argue the applicability of a new method, still in need of validation itself.

A more objective form of validation is presented by Baccino and colleagues (1999) who do not advance their own agenda. They simply present their findings, which consist of a sample of 19 specimens aged seven different ways. They estimate age based on rib and pubic symphysis morphology, dental root translucency, and femoral cortical remodeling. The other three estimates are calculated from a combination of these four methods. They average all the individual age estimates and alternatively estimate age based on an evaluation of the whole skeleton. Lastly, they develop a "two-step" method that estimates young individuals based on pubic morphology and older individuals based on their root translucency. For each of these age estimation methods, the authors present and discuss interobserver agreement and levels of bias and accuracy. Their results give the reader a measure of reliability of the methods of age estimation that they employed.

Osborne and colleagues (2004) looked at the phase method outlined by Lovejoy et al. (1985b) for determining age based on the morphology of the auricular surface. They confirmed that the phase as a whole estimated ages better than either small portions or individual features of the auricular surface. They provided additional positives for the method by showing that sex, ancestry, and the sample source did not significantly influence the outcome of the estimates. Only age itself was tied to the estimation. However, they also found that age only explained 34% of the variation seen in the age indicator. Further negative findings for the Lovejoy phase method were high levels of over aging for both young and old individuals and a failure of the method to place individuals into the correct five-year interval outlined in the methodology

Jackes (1985) presented a validation that was more focused on the methodology and general results of the methods rather than specific measure of performance. The study found that the McKern and Stewart (1957) and the Gilbert and McKern (1973) component methods perform

better on young males and old females. It also found that there was a preponderance of individuals being aged at 35 years, which Jackes found to result from using the mean values as estimates for age. Based on the methodology the author found that the McKern and Stewart method placed too many individuals in the 25 to 45 year age range (Jackes 1985). In this case, rather than evaluate a sample and present the data, Jackes used the validation data and went a step further in explaining what causes the validation results. Jackes actually presented a sort of theoretical validation instead of the usual method validation.

Falys and colleagues (2006) performed a similar validation of the Buckberry and Chamberlain (2002) component system for age determination of the auricular surface. For their data validation they found small amounts of bias with the method but substantial levels of inaccuracy (9.8 years on average). Their study showed better results for higher composite scores. The authors pointed out that combining composite scores into phases obscured variation and showed overoptimistic results for the method. They also found a wide range in ages for particular traits and composite scores and inconsistent progression of age for composite scores.

Similar results were reported by Bedford and colleagues (Bedford et al. 1993), who measured validation performance with accuracy, bias, and observer error. They found that the summary age method of Lovejoy and colleagues (1985a) which combines multiple age indicator ages into one estimate produced better estimates than individual single age indicator methods. Bedford et al. observed regression to the mean for their sample and found that the pubic symphysis worked better for young individuals while the auricular surface provided more accurate age estimates for individuals over the age of 40.

Hens and colleagues (2008) examined both the Suchey-Brooks phase method for age estimation from the pubic symphysis and the Lovejoy method for the auricular surface. The

authors found the same pattern of regression to the mean for both methods where the younger individuals were overestimated and the older ones underestimated. They also found that both methods increased in accuracy with age. Overall in terms of validation statistics (accuracy and bias), the authors found that the Lovejoy method performed slightly better than the Suchey-Brooks method and that the methods worked better for males than females.

Similarly, Millán et al. (2013) examined the Suchey-Brooks phase methods and the Buckberry and Chamberlain component method on a Spanish sample. They found that the two methods performed similarly, but that the method based on the auricular surface performed better over the whole sample. They concluded that the Suchey-Brooks method is more appropriate for younger individuals, while the Buckberry and Chamberlain method performed better on older individuals.

Saunders and colleagues (1992) tested the performance of several different age estimation methods using the common four macroscopic age indicators: pubic symphysis, auricular surface, sternal rib ends, and cranial suture closures. For all their methods they found increasing levels of underestimation of older individuals and overestimation for the young using cranial sutures as well as rib ends. In addition, they found that the auricular surface method had lower levels of correct classification due to the smaller age categories associated with the phases. The pubic symphysis system on the other hand suffered from inherently high inaccuracy due to the very large age ranges associated with its age indicator categories. Saunders and colleagues (1992) found that for both the cranial sutures and rib ends, the methods were only accurate for narrow age ranges, 30 to 56 years and 40 to 59 respectively. The authors also tested the performance of combining multiple age estimates and found that the summary age method was only slightly better than simply averaging all the estimates together. Based on their results the

authors actually advocated *against* the use of multi-trait methods. Like most other anthropologists (i. e., Baccino et al. 1999; Brooks and Suchey 1990; Brooks 1955; Igarashi et al. 2005; Konigsberg et al. 2008; McKern and Stewart 1957; Meindl et al. 1983; Meindl et al. 1985; Todd 1920), they still believe that multiple age indicators should be used to estimate age, but emphasize that individual age estimates should be combined intuitively through experience rather than with a statistical method.

The standard version of a validation study performed on a sample of Thai population is presented by Schmitt (2004). In this study, the Suchey-Brooks phase method and the auricular surface phase method proposed by Lovejoy are tested on an unrelated sample to measure performance. The author presents the results of the study in the form of average biases and inaccuracies for ten year age intervals. Schmitt found that the Suchey-Brooks method tended to underage and bias and inaccuracy increased with chronological age. The study also found that the Suchey-Brooks method estimates ages better for males than females. Lastly, for the sample used in the study only about one third of males and females ended up with age estimates one standard deviation from their actual ages. The use of Lovejoy's method for the auricular surface yielded similar results. The method again tended to underage adults and inaccuracies increased with age. It worked better for males and only about seven percent of specimens were aged into the correct five-year interval the method describes. Therefore, Schmitt (2004) concluded that neither the Suchey-Brooks nor the Lovejoy phase methods perform well on samples from Asian populations. The results of this method show some of the most common results that manifest when anthropologists validate established methods. Frequently the methods do not perform to expectations for the sample being examined, underscoring the importance of validation.

Validation studies provide results that reflect the performance of a methodology more accurately than the results based on the reference sample and the original evaluation of the method.

These are just some examples of validation studies pertaining to age estimation. There are a slew of others and many of them echo the findings outlined in these examples. Some studies obtain results that shed doubt on the valid use of particular age estimation methods while other studies find supporting evidence for the use of certain age estimation methods. It is possible to find support or dissent for any aging method out there. As a scientist attempting to estimate the age of a skeleton, it is often incredibly difficult to know which methods to use. Most colleagues would agree that the more age estimation methods and age indicators are used the better the results should be. Mathematically speaking, it is nevertheless prudent to only use each age indicator once. If one were to adhere to this restriction, it becomes necessary to pick a particular method for a particular age indicator. For myself (and probably a large portion of fellow anthropologists), these methods have been the phase methods introduced for the pubic symphysis (Brooks and Suchey 1990), the auricular surface (Lovejoy et al. 1985b), the sternal rib ends (Işçan et al. 1984a; Işçan et al. 1985), and composite scores for the lateral-anterior and cranial vault suture closures (Masset 1989; Meindl and Lovejoy 1985). These methods have been validated with positive and negative results. They have been around for decades, yet in all this time no new method has come along and become prominent in the mainstream anthropological circles.

Chapter 3: Materials

The Skeletons

The materials utilized for this study were obtained from the William M. Bass Donated Skeletal Collection housed and maintained in the Department of Anthropology at the University of Tennessee, Knoxville. The collection is composed of modern skeletal donations gifted to the university's anthropology program by the individuals themselves, their families, or the medical examiner in cases where no family is available to deal with the remains. Donations primarily come from eastern Tennessee, but several bodies have been donated from other areas of the United States as well. The University of Tennessee has been accepting donations and growing its skeletal collection since 1981. Currently averaging over 100 donations yearly, as of 2015 the collection houses more than 1100 skeletons and is continuing to grow.

Upon an individual's death, the Forensic Anthropology Center (FAC) at the University is contacted. Once the donation paperwork is completed and verified, the bodies are transported to the Anthropological Research Facility (ARF) where they are left to decompose. After decomposition has progressed far enough, the skeletal remains are collected from the ARF and brought to the processing center. Remains are soaked in warm water to loosen connective tissues and subsequently cleaned with toothbrushes and dental picks under warm running water. After the remains are allowed to dry, they are inventoried, measured, and accessioned into the collection where they are individually stored in acid free cardboard boxes.

Demographically, the Bass Donated Collection is primarily male, white, and is composed of elderly individuals, as the majority of donors die of natural causes associated with advanced age. The collection's demography as a whole may not represent the ideal collection for age

estimation studies. Most age estimation methods focus on the younger age ranges (between 20 and 50 years), and the Bass Donated Collection is definitely top-heavy in its age distribution. It also does not present a heterogeneous cross section of the modern American population. However, the collection is very clean, very complete, and well maintained. Potentially its best attribute is that, since the majority of individuals have filled out extensive information forms as part of the donation process, the exact ages are known for the skeletons in the collection. Some other older collections have problems in that out of thousands of skeletons, the ages of only a couple of hundred are known with any degree of certainty (Lovejoy et al. 1985a).

The focus of this study lies in the performance of existing age indicators and the introduction of a new approach for building age estimation techniques. As the establishment of a well-founded, accurate, and replicable age estimation method was not the primary goal here, the representativeness of the collection's demography was deemed of lesser importance than the completeness of the skeletons—both in terms of inventory and preservation—and the reliability of ages on record.

The Sample

One hundred seventy-six skeletons were selected from the William M. Bass Donated Collection. Sex and ancestry were disregarded in the selection process. Individuals were selected on the basis of their age and the completeness of their skeleton. As the study required looking at multiple age indicators throughout the body, the skeletons selected were preferred to have good representation of elements. In order to collect data on all age indicators used in the study, each skeleton needed to have one pubic symphysis and one auricular surface in good condition. It also needed to have one identifiable fourth rib, which usually required several other ribs in order to

securely identify the fourth. Lastly, over half of the cranium needed to be intact to allow scoring of midline sutures and at least one side of the bilateral sutures. While these were the minimum requirements, efforts were made to select specimens with as complete a representation of the skeletal elements of interest as possible. These elements of interest were the os coxae with both the symphysis and auricular surface observable and undamaged, the intact and complete cranium, and as many identifiable of six sternal rib ends as possible (the left and right third, fourth, and fifth ribs).

In addition to completeness, individual specimens were also selected based on age. Ideally, age estimation methods should be applicable to be used on any skeleton. However, there are several reasons anthropologists are aware of the limits of aging methods to approximate age accurately outside the range of the original test sample. The first reason is mathematical. Extrapolation—estimation outside the known range—grows increasingly more inaccurate the further from the known range you want to estimate. Secondly, the age indicator definitions are based on what is observable; individuals outside the observed age range may exhibit characteristics that were not included in the age indicator descriptions. Here it was illogical to include old individuals as the established age estimation methods are generally not applicable to these ages and would have severely skewed the results. Furthermore, few of the age estimation techniques used herein have been shown to accurately work on individuals over the age of 55. It would have been redundant to collect data on elderly individuals and then conclude that age estimation methods were not appropriate to be used on these specimens. To keep the methods viable, this study did not select any individuals over the age of 60.

In terms of age ranges, individuals were selected in such a way as to be equally represented in the data. Approximately 30 individuals were selected for every five-year interval.

For those intervals that did not have 30 individuals, as many individuals as were available that met the selection criteria were utilized. All specimens were selected during the days preceding the data collection. On selection days, the completeness and appropriateness of each skeleton was checked and all demographic information on the skeleton's storage box was blocked to prevent it from being read. In this way, the researcher hoped to eliminate any bias during the data collection, so that at no point would there be knowledge of the identity, especially the age, of the specimen being examined. Table 3.1 shows the demographic distribution of the sample based on recorded sex and ancestry. Table 3.2 gives basic measures of the age distribution of the sample used, and Table 3.3 shows a breakdown of the study sample by age groups. Figure 3.1 shows the age distribution in histogram form.

Table 3.1: Demographic distribution of individuals in the study sample

	White	Black	Hispanic	Native Am.	Total
Male	124	19	5	1	149
Female	24	3	0	0	27
Total	148	22	5	1	176

Table 3.2: Basic statistics of the age distribution of the study sample

Mean	St. Dev	Min	Max
45.48	9.495	23	60

Table 3.3: Distribution of Individuals in the study sample based on age groups

Age Group	Male	Female	Total
20-25	4	1	5
26-30	7	0	7
31-35	13	3	16
36-40	18	7	25
41-45	26	4	30
46-50	29	3	32
51-55	27	3	30
56-60	25	6	31
Total	149	27	176

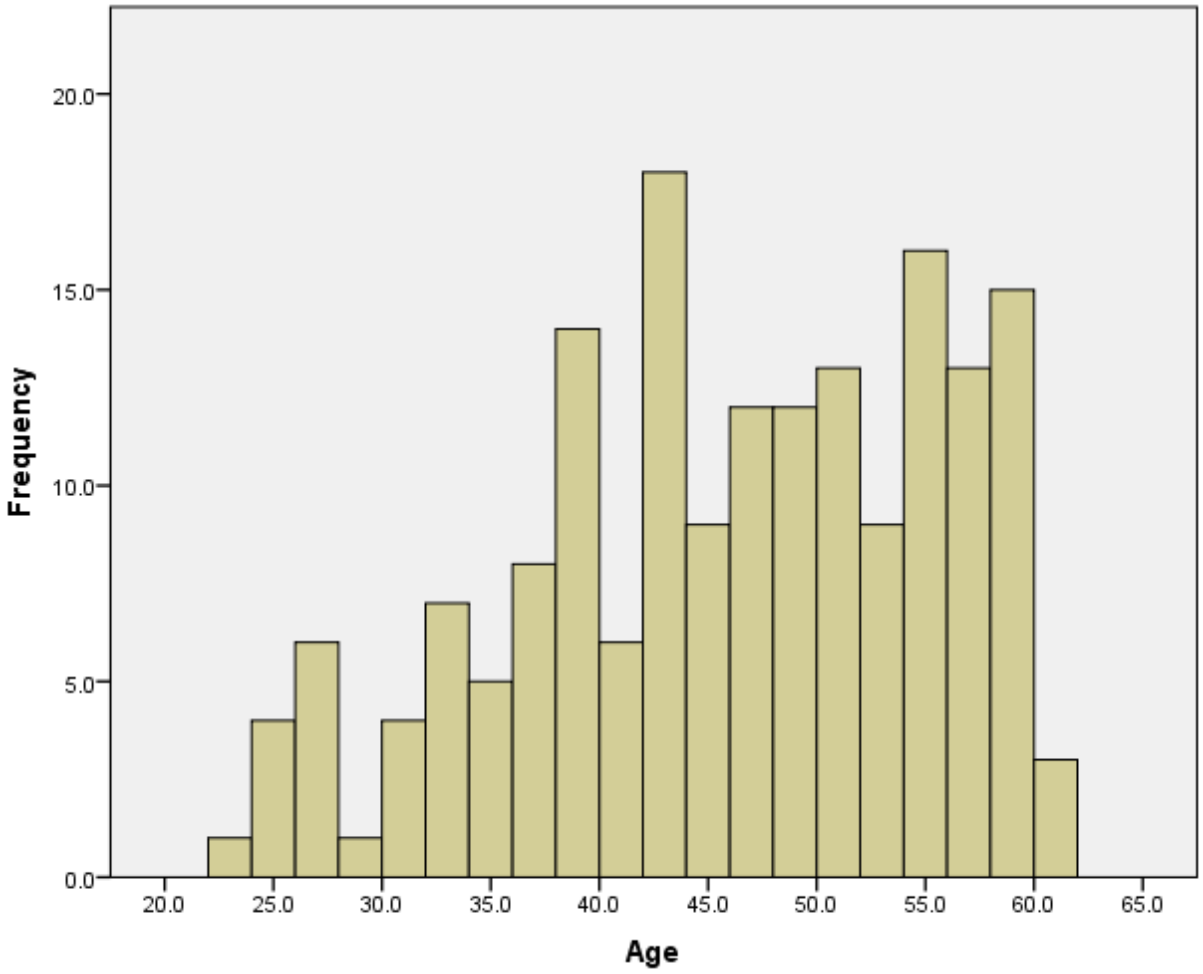


Figure 3.1: Histogram of the age distribution of the study sample

The Variables

In total, 188 data observations were recorded for each of the 176 skeletons in this study. Of the 188, 10 were midline variables that were only recorded once; the remaining 178 were bilateral observations of 89 variables recorded on both sides. All 10 midline variables as well as 82 bilateral variables were taken from the literature. Some of the variable descriptions in the literature were very confusing, making it difficult to identify exactly what the authors were referring to. Still other variables appeared to have very little applicability to the data under study (i.e., most observations were the maximum stage in the data; as such the changes defined in the variable were not observable in this data because they occurred very early.) For these reasons, seven additional variables were defined by the author to clarify unclear, ambiguous, or useless (for this sample) literary definitions. The literary definitions were kept as well and applied as consistently as possible during data collection.

For the pubic symphysis, phases were recorded based on the descriptions of Todd (1920; 1921) and those outlined by Brooks and Suchey (1990). Many different components, outlined below, were recorded for the morphology of the pubic symphysis. Hanihara (1978) defined seven components (horizontal ridges, pubic tubercle, lower end, dorsal margin, superior ossific nodule, ventral beveling, and symphyseal rim). McKern and Stewart defined three components (dorsal demiface, ventral rampart, and symphyseal rim) in 1957 and Gilbert and McKern redefined the same three components based on a female sample (1973). Chen and colleagues (2008) defined nine different components (ridges and furrows, ridges of the pubic tubercle, lower extremity, ventral rampart, ossific nodule, dorsal margin, ventral beveling, general macroscopic changes, and bone density). Lastly, Boldsen and colleagues listed descriptions for five pubic

symphysis components in 2002 (symphyseal relief, symphyseal texture, superior apex, ventral symphyseal margin, and dorsal symphyseal margin).

For the auricular surface, the phase descriptions of Lovejoy and colleagues (1985b) were used. Nine component descriptions were taken from Boldsen et al. (2002) (superior demiface topography, inferior demiface topography, superior surface morphology, apical surface morphology, inferior surface morphology, inferior surface texture, superior posterior iliac exostoses, inferior posterior iliac exostoses, and posterior iliac exostoses). Buckberry and Chamberlain provided definitions for five components in their article from 2002 (transverse organization, surface texture, microporosity, macroporosity, apical changes). Thirteen further binary components were taken from the definitions given by Igarashi and colleagues in 2005 (wide groove, striation, roughness, flatness, smoothness, fine granularity, coarse granularity, sparse porosity, dense porosity, dull rim, lipping, tuberosity, and bony bridge).

For the morphology of the sternal rib ends, the phase descriptions provided by Isçan and colleagues (1984a, 1985) were used. Three rib end components defined by Isçan et al. were also used (pit depth, pit shape, rim and wall configuration) (1984b).

Cranial suture closure definitions were taken from two sources. Meindl and Lovejoy's (1985) definitions were used to score six bilateral and eight midline and maxillary sutures, and the definitions given by Boldsen and colleagues (2002) were used to score three bilateral and two midline sutures.

As no demographic details were known by the observer during data collection, all sex-specific methods (i.e., Todd 1920; Todd 1921; McKern and Stewart 1957; Gilbert and McKern 1973; Isçan et al. 1984a; Isçan et al. 1985) were applied to all skeletons regardless of sex. This means that male specimens were evaluated with methods designed for females and vice versa.

See Appendix A for detailed tables and lists of all variables used in this study. The tables in Appendix A also provide the variable codes that will appear in the output and the number of stages for which each variable was defined. Appendix B provides the stage descriptions and definitions from the literature used in this study. Appendix A also provides information on the author-defined variables, from which variables they were derived and their codes. Their descriptions are given in Appendix B under the method from which they were derived.

Chapter 4: Methods

Variable Manipulation

Missing Data

While care was taken to select skeletons with observable age indicators, not all variables were observable for all of the specimens in the study sample. Missing bilateral data were replaced using the opposite side. While asymmetry is present in the sample, the opposite side still provides the best approximation for a missing variable, and since only one side was used in the analysis, no duplicated values were used. Missing values for the age indicators of the ribs were a little bit more complicated. The primary rib of interest was the fourth rib. Analysis of the differences between the fourth and the adjacent third and fifth ribs showed that the fifth was more frequently in agreement with the fourth rib than the third. Therefore, a missing variable for a fourth rib was firstly approximated from the opposite fourth rib. When neither of the fourth ribs was observable, the fifth ribs were used as proxies, with the same side fifth rib given priority over the opposite side. When neither of the fourth or fifth ribs was observable, the missing value for the fourth rib was approximated from the value of the third rib of the same side, leaving the value of the third rib's age indicator of the opposite side as the final approximation for the value of the missing fourth rib.

These substitutions filled in most of the missing values present in the data set. The analyses did not utilize variables from the left and the right side at the same time, thus none of the data values duplicated by the substitution process were used more than once in any of the statistical models. Likewise, the data for the third and fifth ribs were not used in these analyses, eliminating duplicated substitution values from being used multiple times in the same analysis

and causing artificial homogeneity in the data. In total 89 variables were used from the literature in the analysis.

Some midline variables were not observable and for some of the specimens neither of the sides could be recorded. Normally, these kinds of missing values would have to be approximated from other variables in the data set; however, as the number of these kinds of missing values was relatively low, approximation of these missing values using regression or correlation analyses was not deemed necessary in this study. Overall, the amount of missing data values after the substitutions outlined above was minor.

Additional Variables

Five additional variables were created for those sets of methods that were sex-specific (i.e., Todd 1920; Todd 1921; McKern and Stewart 1957; Gilbert and McKern 1973; Isçan et al. 1984a; Isçan et al. 1985). Without knowing the sex of the individual being aged, each specimen was aged using both the male and female standards. Later, when the sex of the individual was known, the specimens were re-evaluated according to the correct method based on sex. This resulted in 5 new variables. Variable M1.5 represents the sex-specific pubic phase score (Todd 1920; Todd 1921). Variable M12.5 represents the sex-specific phase method for the sternal rib ends (Isçan et al. 1984a; Isçan et al. 1985). The three variables M5.5 C1, M5.5 C2, and M5.5 C3 represent the sex-specific scores for the three components of the McKern methodologies--dorsal demiface, ventral rampart, and symphyseal rim, respectively (McKern and Stewart 1957; Gilbert and McKern 1973). With the addition of these five variables to the 79 left side variables and the 10 midline variables, a total of 94 variables were used in the analysis. These variables are outlined in Table A.6.

Decision Trees

Analysis of the data was performed using SPSS 22. Only the data from the left side with the substitutions outlined in the previous section were used for the analysis of the data. The data were analyzed using decision trees, classification models that use splitting values to separate a data set into similar smaller groups. They are called decision "trees" precisely because when displayed graphically, they depict a branching system that looks like a stylized root system, a family tree, or a flow chart. They use different statistical calculations, or growing methods, to calculate values that efficiently split the data set into smaller groups, which are split further and further until a growth stopping condition is met. Figure 4.1 illustrates how a decision tree works and should be read, using an example of a basic decision tree diagram with decisions based on the topic of this study. The decisions depicted in this example are solely for explanatory purposes and are not based on any statistical analysis.

Based on the depiction in Figure 4.1 it is possible to talk about some basic terminology and use of decision trees. In a decision tree, the data is split into groups called nodes. The first such node contains the complete data set and is called the "root node" (Node 0). Each step in the decision tree analysis splits the data into two or more new nodes that contain a section of the data from the original node. In this way a "parent" node (e.g. Node 0) gives rise to several—in this case, 2—"child" nodes (Nodes 1 and 2). The observations in each child node are distinct; when the observations out of two child nodes are combined, the combination would represent the data that was in the parent node. With further splits, each child node can in turn become a parent node giving rise to its own child nodes. In Figure 4.1, Node 1 becomes parent to Nodes 3 and 4 and Node 2 gives rise to Nodes 5 and 6.

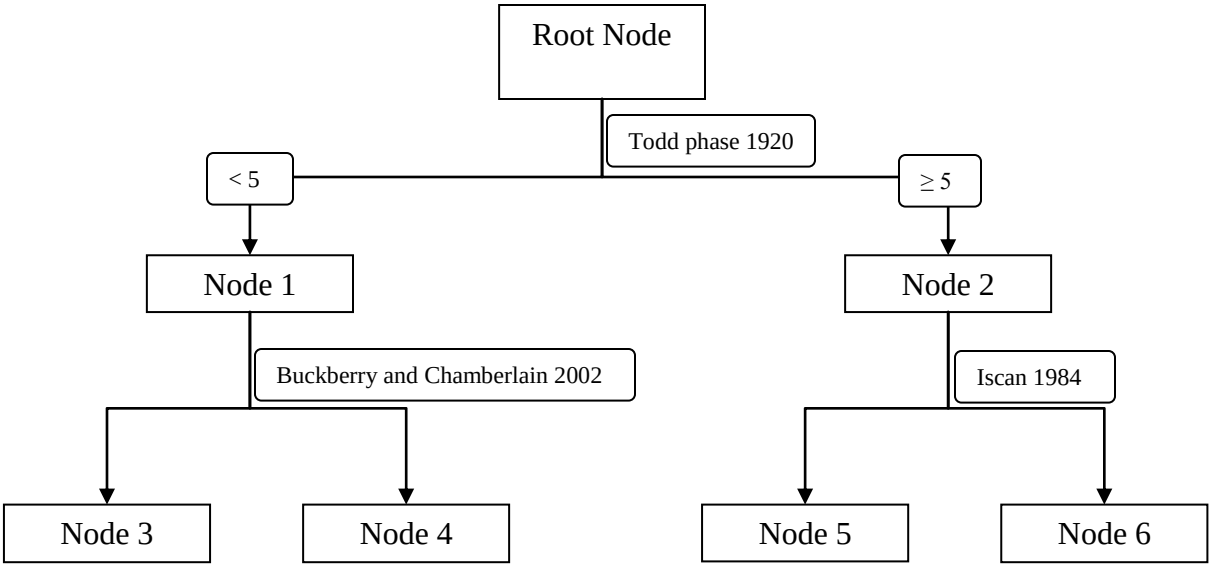


Figure 4.1: Basic decision tree with example decisions

In the example above, some fictitious decisions are included to illustrate the use and function of a decision tree as it might pertain to the issue of age estimation. As mentioned earlier, the root node contains all the data. This particular decision tree algorithm determined that of all the variables used, the most significant and efficient split for the data was based on the pubic symphysis phase method of Todd (1920). The split in the method is between pubic symphysis phases 4 and 5, therefore, the root node is split into two child nodes based on the value of the Todd phase of the pubic symphysis. If the symphysis phase is less than five (i.e., having a value of four or less), the observation goes into node 1. If the symphysis phase is five or greater, the observation goes into node 2. In this case, the second level of the decision tree would have two nodes, one (Node 1) containing all observations with a Todd phase of less than five and the other (Node 2) containing the observations with a Todd phase of greater than or equal to five. The observations in Node 1 are split further by the condition of the apical score based on the definitions of Buckberry and Chamberlain (2002). Node 2 on the other hand is split by a

different condition, in this case, the Iscan rib phases (1984a). Node 2 would contain the older specimens in the data sample and Node 1 the younger ones, based on the notion that larger Todd phases belong to older individuals. Furthermore, this particular decision tree would indicate that the Iscan rib phases are more useful to separate the older individuals (from Node 2), while the younger individuals are more easily distinguished by the Buckberry apical score. This simple example should illustrate the potential benefit decision tree might offer to age estimation.

Decision trees are commonly cited as examples of how banks determine high- and low-risk candidates for loans. In these examples, decision trees are used to come up with a rule set to determine who is a good candidate to receive a loan. Decision trees are used to classify cases into one of several categories, such as low-, medium-, or high-risk. Besides determining group classification, decision trees can also be used to create predictive models for new cases. Ultimately, this predictive modeling would be the reason decision trees could be utilized for age estimation (IBM SPSS Statistics 2013).

First, however, decision trees are useful for a different reason. The data set to be analyzed in this study contains 94 different age indicator variables, not including any right-sided variables. The primary interests of the study is to narrow that field of variables down to those that are significant and useful in age estimation. Decision tree analyses can be used to data mine a large number of variables and find significant or useful variables or subsets of variables for analysis. Decision trees can also highlight interactions between the variables and between the data and the variables. A decision tree can also show that a significant relationship between the variables may only exist in some subgroups of the data. A decision tree does not have to apply all the significant variables to the whole data set. As this study is more interested in the age indicator

variables than in the age estimation power of an aging method, both the variable screening and interaction identification properties of decision tree analysis make it an applicable method.

Further benefits of decision trees include the output they produce and the requirements needed for the analysis. As shown in the example above, the visual output of a decision tree is quite straightforward and intuitive. Decision tree diagrams are easy to understand even by anthropologists without a strong statistical background. On the requirement front, decision trees have no problem handling ordinal and nominal variables in addition to continuous variables and can even be grown to account for missing data. Given that osteological age indicators are primarily descriptive (i.e., categorical) in nature with an implied direction to them (i.e. ordinal), decision trees seem like a very good fit to analyze a large set of age indicators variables.

In this study almost all variables were treated as ordinal. Only age and the angle of the ventral pubic angle (M7 Ga L) were treated as continuous variables.

Analysis

Decision trees were constructed using two growing methods: the chi-squared automatic interaction detection (CHAID) and the classification and regression tree (CRT) method. CHAID chooses the independent variable with the strongest interaction with the target variable at each growth step in the decision tree formation. This method uses variable relationships as the determining factor behind growing the decision tree. CHAID also has the benefit of being able to create multi-way splits, where a parent node gives rise to more than just two child nodes. CRT separates the data into groups of observations that are as similar to each other as possible. In other words, CRT attempts to minimize within-tree node variability or homogenize each node. While CHAID may produce multi-way splits, CRT always creates two child nodes from a parent

node (a binary split); however, CRT has the added benefit of allowing the grown tree to be pruned.

Once the complete tree is calculated and has stopped growing, pruning can be performed to reduce the size of the decision tree. The pruning procedure eliminates different splits in the tree and checks if that simpler tree is significantly different from the original. It tests these smaller subtrees and finds the simplest one that falls within a specified level of performance of the original decision tree. Generally, pruning ends up eliminating the least important variable or variables in the model, making the model slightly simpler. It does not change the core structure of the tree, nor does it change the variables selected for the decision tree. It only removes a couple unimportant ones.

The process of pruning a decision tree generally involves removing the variables in the tree with the lowest influence or significance. However, since the decision tree chose the variables from the data set, they may still provide some useful information—information that would be lost if pruned from the tree. As the primary focus of the study presented in this paper was to determine which variables would be selected by the decision tree method, and not the efficiency of the decision tree model, pruning was not performed on the data.

As mentioned previously, decision trees can account for missing data, however, the two growing methods handle missing data differently. In the CHAID growing method, missing data are treated as a new category that is added to the model. For example, a binary variable with missing values would be used in the analysis as if it had three possible values (i.e., A, B, and Missing) instead of two. Due to the low number of missing values in the data used for this study, it is unlikely that the missing values would have a noticeable effect on the CHAID decision trees grown. In addition, the use of missing data would serve well as an example of further

application for decision trees in the field of anthropology where missingness can be a common problem.

In contrast to the CHAID, the CRT growing method uses surrogates to fill in missing data. This means that whenever a split occurs in the tree and the decision variable contains missing data, the CRT method finds a variable closely related to the missing one and uses the surrogate variable to determine on which side of the split the observation with the missing value belongs. It is also possible to disallow surrogate variables, in which case the observations with missing values are simply excluded from the model. In this study, the author opted not to use surrogates in order to avoid obscuring the results and to allow for easier interpretation of the decision trees.

The purpose of this study is to identify which age indicators and age indicator variables are useful for age estimation. To this end, several decision trees were grown using both the CRT and CHAID growth methods. Several trees were constructed using each method based on a different set of parameters. For each of the resulting trees, the variables that were chosen were noted. After viewing the variables chosen for all the different decision trees, the most influential age indicator variables were determined. While predictive performance of the decision trees was not the focus of this study, it was still tested in order to compare the models to each other and allow evaluation of which models had achieved better age estimation. The performance of each decision tree was calculated in two main numbers. First, for each model the prediction errors were calculated and the absolute values were averaged to arrive at the mean absolute error (MAE). Second, the percentage of the variance in age accounted for by the model was calculated by the following equation:

$$\% \, Var = 1 - \frac{\sum(observed - predicted)^2/n}{s^2}$$

where the numerator is the sum of squared residuals of all observations divided by the number of observations in the decision tree (n) and the denominator is the squared standard deviation in the root node of the decision tree, which is the variance of age for the observations used in the decision tree.

When starting a decision tree, one of the options able to be manipulated is the desired node size. Two parameters can be set: the minimum parent node size and the minimum child node size. Both of these parameters set a limit on the number of splits that are found during the growing process. The minimum parent node size gives the number of observations that have to be in a node for it to be considered split-worthy. Hence, if a node is smaller than the minimum parent node size, the method will not attempt to split the node into smaller nodes. The minimum child node size sets a limit on which splits are to be considered worthwhile. If the methodology finds a significant split in the data but one or both of the child nodes would have less than the minimum child node size number of observations, that split is not to be considered necessary.

Both of these parameters limit the number of splits and nodes that are determined in the decision tree procedure and smaller numbers allow the tree to grow larger. These numbers are heavily dependent on the sample size of the data being used. The data sample of this study has a size of 176, and, with the parent/child node parameters set to 100/50, the resulting decision trees are very small. With the data set used in this study, the CRT procedure actually fails to split the data at all, because all child nodes would end up being smaller than 50 observations in size. Decision trees are often used on very large data sets—compared to small anthropological sample sizes—that would not run into this particular problem. Due to the relatively small sample size of

this study the parent and child node size parameters were adjusted to allow the decision tree to grow to a larger size that allowed for more interpretation.

Initially, values of 10 and 5 were used to show the "maximum" size of the decision tree. These trees were huge and extremely complex. The size of these trees made interpretation of the patterns shown in the variables difficult. Further, the complexity of these trees likely represented an overfitting of the model to the data. This means that the model was too specifically tailored to the data and any interpretations drawn from the tree would be more indicative of the data under study than age estimation as a whole. In order to produce simpler decision trees that would provide more reliable interpretations, the parameters for parent and child minimum node sizes were changed. Various sets were tested until a good combination was determined. The values that were found to work well were 30 for parent node size and 10 for child node size. The decision trees resulting from the use of these parameters were simpler, easier to understand, and allowed interpretation of the relationships between the age indicator variables selected and age.

With the parent and child node sizes set, several decision tree models were constructed using both the CRT and CHAID growing methods. The first trees were grown using the complete set of 94 left and midline variables. Residuals were calculated and plotted against predicted age. If they were flagged as extreme they were removed from the data set and the decision tree was reconstructed. This process was repeated until either no more residuals showed up or removing the residual outliers did not alter the structure of the decision tree.

For purposes of comparison, decision trees were also constructed on different subsets of the age indicator variables; specifically, using only variables from age determination methods based on phase descriptions. To complement these models further, decision trees were constructed using only the variables from age determination methods based on component age

indicators. Finally, a CRT and CHAID tree was constructed using only the methods to which this researcher had been exposed and taught to use for age estimation. These methods were the pubic symphysis phase methods of Todd (1920; 1921) and Suchey-Brooks (1990); the pubic component methods of McKern and Stewart (1957) and Gilbert and McKern (1973); the phase method of the auricular surface (Lovejoy et al. 1985b); and the phase method of the sternal rib end morphology (Iscan et al 1984a; Iscan et al. 1985).

After the construction of each of the decision trees, the variables selected by each of the tree models were used in the construction of a regression equation that would allow comparison of the decision tree results to a better-known statistical method. For each regression equation, the mean absolute prediction error (MAE) was calculated. The R^2 values for the regression equations were noted as they provide comparable numbers to the percentage of the data variance explained that was calculated for each of the decision trees. Lastly, a stepwise variable selection regression analysis was performed to see which of the 94 variables a regression analysis would find to be the most informative with respect to the biological age - age indicator relationship. The models were compared both on the basis of the proportion of the variance explained by the model as well as the mean absolute error.

In order to determine the most influential variables in the age estimation procedure, the variables that appeared in each distinct model were noted. The number of models in which each variable appeared was calculated and the most frequently appearing variables were noted and judged to be the most important variables for age estimation, or at least have important information to provide in an age estimation process.

In addition to analyses performed on the complete data set, several further analyses were run on subsets of the data. This was done to see if the same variables were important for all the

observations, or if there were some variables that would work better on some portion of the age distribution that other portions. To accomplish this, the data was divided into six distinct age groups of similar size. Each age group was constituted of a five-year age interval. However, due to the limited sample size for the young ages in the sample, the youngest three five-year intervals were combined into a single age interval. Table 4.1 gives the age ranges of the six age groups and their corresponding sample sizes

In order to investigate the influence of different age groups on the variables selected and the decision trees grown, these age groups were tested in a jackknife fashion. This means that six different subsets of the data were analyzed. For each of the subsets one of the age groups was held out and models were constructed using the rest of the data. For each jackknife data set a CHAID 30/10 and CRT 30/10 decision tree were constructed. For these trees all 94 variables were entered into the analysis, the primary interest lying in how the decision trees change as successive age ranges were held out of the decision tree construction.

Table 4.1: Breakdown of Age Groups

Age Group	Age Range	Sample Size
1	23-35	28
2	36-40	25
3	41-45	30
4	46-50	32
5	51-55	30
6	56-60	31

Additionally, each of these jackknife trees was constructed using 10-sample folds. This 10-fold cross-validation means that for each decision tree, the data set was divided into ten subsamples and then ten decision trees are constructed with each of the subsample folds held back for one tree. In the end, only the tree with the best performance on the fold holdout is selected as the best one of the ten trees constructed. This cross-validation produces more robust models that are less influenced by extreme individual observations. These decision trees were compared on the basis of the variables selected in each and the mean absolute error (MAE) of the observations in the jackknifed age group and the mean absolute error for the observations used to construct the decision tree.

Besides the two decision trees that were constructed for each jackknife data subsample, a stepwise logistic regression model was also constructed with each of the age groups held out. These logistic regressions were compared on the basis of the variables selected for each regression and the proportion of misclassifications of the observations.

Chapter 5: Results

Notation

Before the actual results are presented it would prove useful to outline some of the short hand notation that will be used. The two tree growing methods and their respective abbreviations that will be used were already mentioned: the chi-squared automatic interaction detection (CHAID) and the classification and regression tree (CRT) method. Generally speaking when referring to a particular decision tree, the form will be of the type of tree followed by two numbers separated by a forward slash, such as "CHAID 30/10". The two numbers refer to the node size parameters, the first equaling the minimum parent node size and the second number referring to the minimum child node size. Additional notation will follow the main portion and can refer to a subset of variables (vars) used or observations (obs) omitted from the data analysis. If no additional information is presented, the default should be assumed to apply unless otherwise stated. The default for variables is all 94 midline and left-sided variables and for data the default is all 176 observations. Having said that, the phrase "CHAID 40/20 component vars only without obs 156" should pose no further questions other than "Where can I find this masterwork of statistical computing?"

Some of the notations to the splitting rules can be unfamiliar. Most should be familiar with "greater than" (>), "less than" (<), "greater than or equal to" (≥), and "less than or equal to" (≤). However, the decision tree also uses brackets and parentheses to denote ranges. In this case brackets denote inclusiveness and parentheses exclusiveness. Therefore a range of (3 to 6] means greater than 3 and less than or equal to 6, or everything between 3 and 6 including 6, but not 3.

In this study this notation is only used for integer ranges, so a range of (3 to 6] would mean values 4, 5, and 6.

Furthermore, all of the decision tree diagrams use consistent variable coding for all 94 variables used throughout this analysis. For convenience sake, the most pertinent variables are mentioned in the writing. For those who want to know exactly which were chosen by each and every decision tree, all the variables and their corresponding codes can be found in Appendix A in Tables A.1 through A.6 along with the sources from which the variables were taken.

Chi-Squared Automatic Interaction Detection (CHAID) Decision Trees

The first decision tree that was grown was a default CHAID tree, with all other values and parameters left at their respective default values. That would be a CHAID 100/50 tree and the graphical representation of this decision tree looks like the depiction in Figure 5.1. Below each node is noted the variable used to split the node and along each branch is the classification criteria used to determine the contents of the child nodes. Within each node, data on the observations contained in it are given, including the mean, standard deviation, the number of observations in the node, what percentage of observations of the whole data set are in the node, and the predicted value for the observations in the node. All the values given in the node pertain to age and as such are in terms of years (except for the percent and the size). The most important value in the node is the predicted value, as this is the value that would be used for age estimation based on the tree. An individual's age would be estimated as the mean age of the individuals in the node that the specimen gets sorted into. So for every decision tree the number of terminal nodes is equal to the number of distinct age estimates that decision tree would predict. By

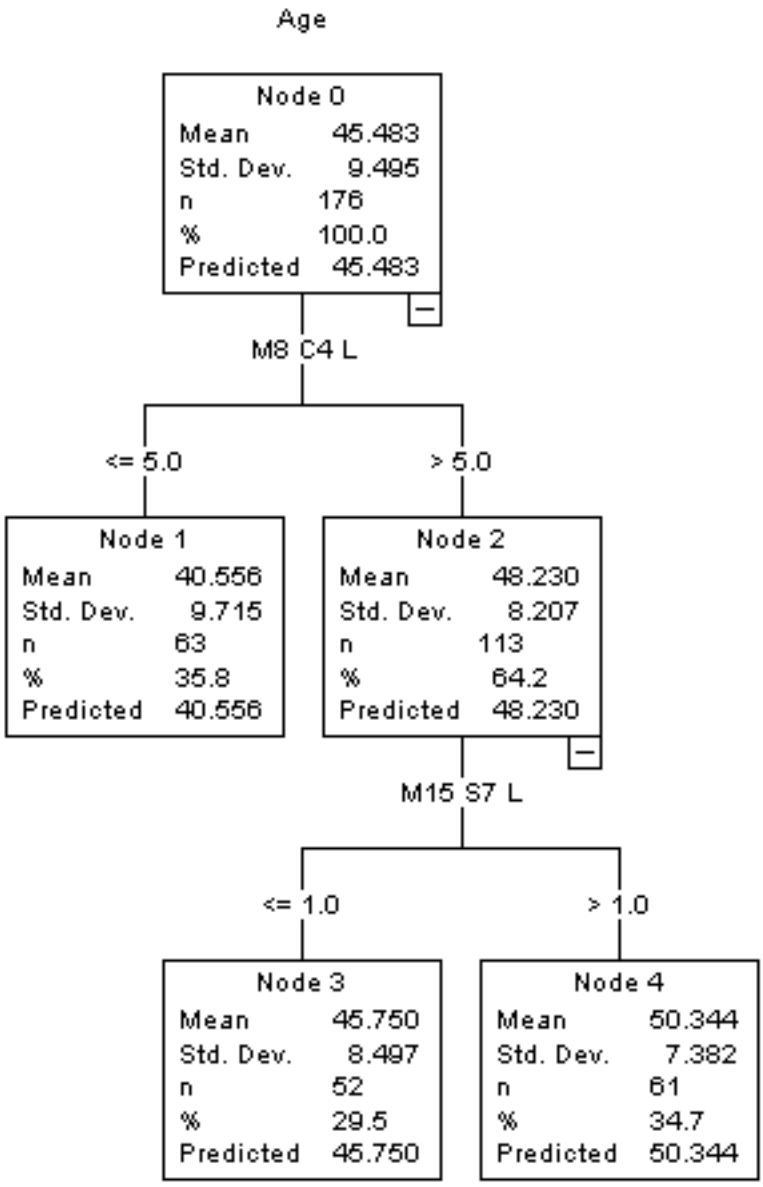


Figure 5.1: CHAID 100/50 tree with all variables, observations, and all default settings

comparison, traditional age estimation methods have about six (Brooks and Suchey 1990) to ten (Todd 1920) different age estimates.

This decision tree separates the data into three groups determined by their scores for the pubic ventral symphyseal margin based on the definitions by Boldsen et al. (2002) and the degree of suture closure at pterion using the definitions of Meindl and Lovejoy (1985). The youngest group of individuals are those with a ventral margin score of less than or equal to 5 and would be estimated to be about 41 years old. The middle group would be those with a ventral margin score of greater than 5 and a pterion suture closure score of 1 or less. These individuals are estimated to be about 46 years old. The oldest group of this decision tree would be those individuals with a ventral margin score of greater than 5 and a pterion suture score of greater than 1 and would be estimated to be 50 years of age. This decision tree is very simple and easy to understand.

The CHAID 100/50 tree is a good example of how the parent and child node size parameters limit tree growth. As none of the three terminal nodes in this tree have more than 100 observations, the procedure will not even consider splitting these nodes further as the minimum parent node size is set to 100. Thus, a large minimum parent node size value limits the growth of the tree and potentially terminates growth prematurely. The child node size similarly limits the growth of a tree.

After the initial "default" CHAID tree, the next decision tree grown was CHAID 10/5 seen in Figure 5.2. It is extensive, complicated, and quite overwhelming. By reducing the values of the minimum node sizes, the tree is allowed to grow to a large complex size as the decision tree algorithm is allowed to continue splitting child nodes further into the tree. This tree is more likely to select variables that are related to age than a tree that is limited by restrictions placed on

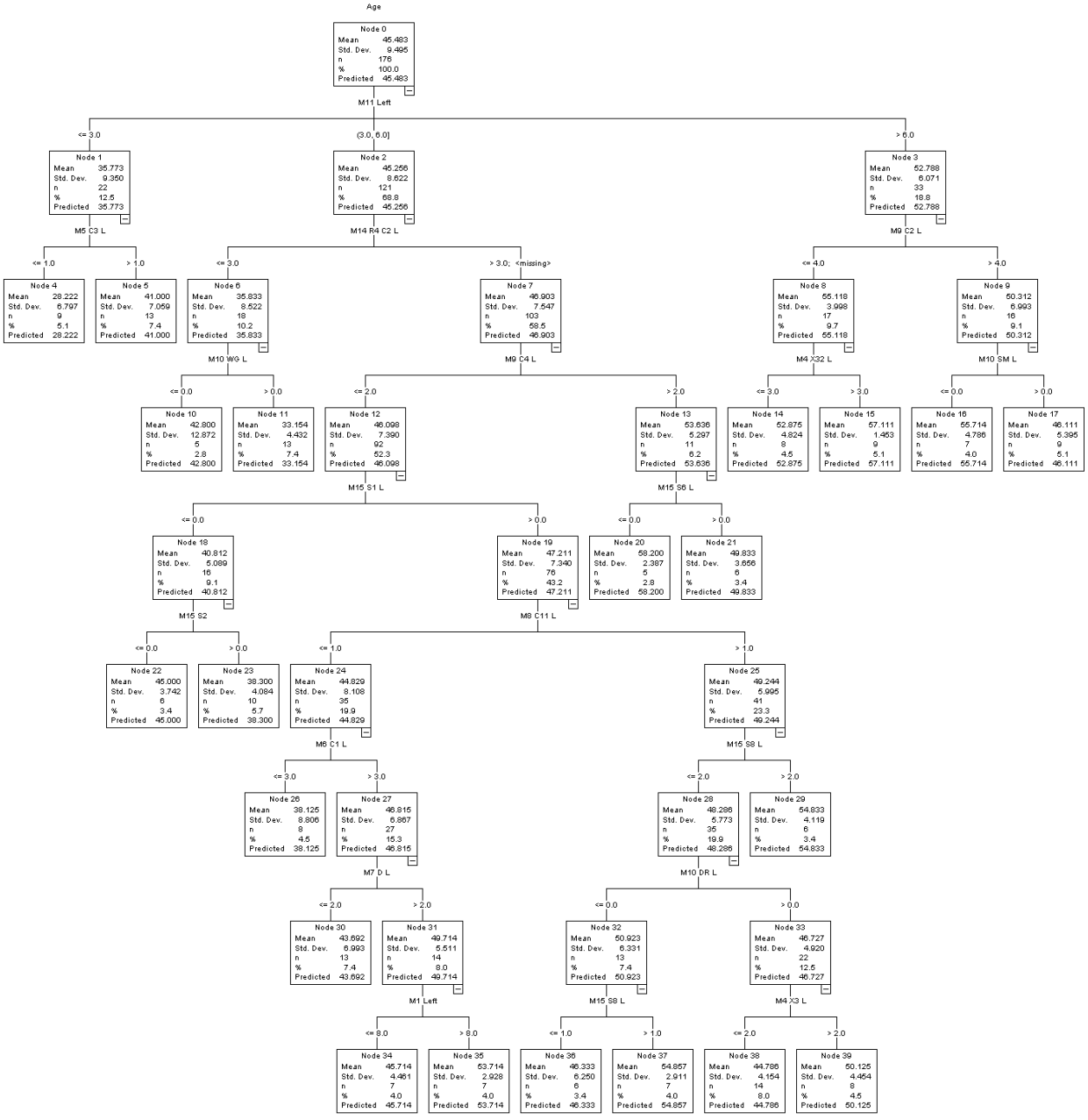


Figure 5.2: CHAID 10/5 with all variables, observations, and otherwise default settings

the tree's node sizes. In the CHAID 10/5 tree, the most important variable is the phase system scoring defined by Lovejoy et al. (1985b) used to split the root node. This initial variable splits the data into three groups. The youngest group has an auricular phase score of less than 4. This group is further split by the score of the symphyseal rim as defined by McKern and Stewart (1957). The oldest group is defined by having an auricular phase score greater than 6. The oldest group is split further by the surface texture of the auricular surface, coded by Buckberry and Chamberlain (2002). The middle group is split along various variables. The central part of the decision tree is formed from variables based on the pit shape of the rib (İşcan et al. 1984b), macroporosity on the auricular surface (Buckberry and Chamberlain 2002), suture closure at midlambdoid (Meindl and Lovejoy 1985) and the surface texture of the inferior portion of the auricular surface (Boldsen et al. 2002). The rest of the variables that are part of the tree define small splits which are not listed here. Overall, the tree uses 18 variables to create 21 age groups. The estimates for these age groups range from 28.2 years to 58.2 years, which covers a good portion of the age range in the sample. Using the terminal nodes to estimate the ages of the observations in the sample yields a mean absolute error of 3.81 years. This decision tree accounts for 70.25% of the variance in age.

Both of these performance values are based on the decision tree being evaluated with the data that was used to create it. As such the values are inflated. They are better than they would be if this decision tree was actually used to estimate ages for individuals from a validation sample. This study does not use a validation sample. As such, all the performance values in this study are inflated, but can still be used to compare the models in this study to each other. With the small node size requirements, the tree is able to grow very large, because the nodes are allowed to be split to smaller sizes. The terminal nodes in CHAID 10/5 are very small, averaging

around 10 observations. With each node being so small, the age estimate for the nodes is tied to a very specific group of observations. While it is possible to claim that each terminal group represents a distinct age group of individuals with similar traits, the small size of the group prevents concluding that this group would still be distinguishable in the population or a different sample. The small node sizes cause the age estimates to be extremely specific to the data on which the model is built and would not be helpful for most data outside of this study. This type of model specificity is called overfitting. Conclusions drawn from an overfitted model cannot be reliably applied to the field as the conclusions and interpretations actually reflect more of what is going on in the study sample than the population. While the complete CHAID 10/5 decision tree should not be used to estimate ages, the relationships between the age indicator variables and age indicated by the variables chosen in the core of the tree are still valuable. Core refers to the major variables that split large nodes. The smaller nodes and the variables associated with them are likely artifacts of the model's overfitting.

Evaluation of the residuals from CHAID 10/5 showed several observations that had very large residuals and could be considered outliers. The two observations with the largest residuals were observations 81 and 116 as can be seen in Figure 5.3. These two observations were removed and a new CHAID 10/5 decision tree was constructed. This new tree is presented in Figure C.1 in the appendix. This new decision tree is identical in structure to the original CHAID 10/5 tree. The initial branching is still determined by the evaluation of the Lovejoy auricular surface phase (1985b). Since this tree is the same as the previous one, removal of the two large residuals from CHAID 10/5 did not affect the structure of the decision tree. This indicates that while these two observations may have had very incorrect estimated ages, they did not have enough influence by themselves to change the decision tree.

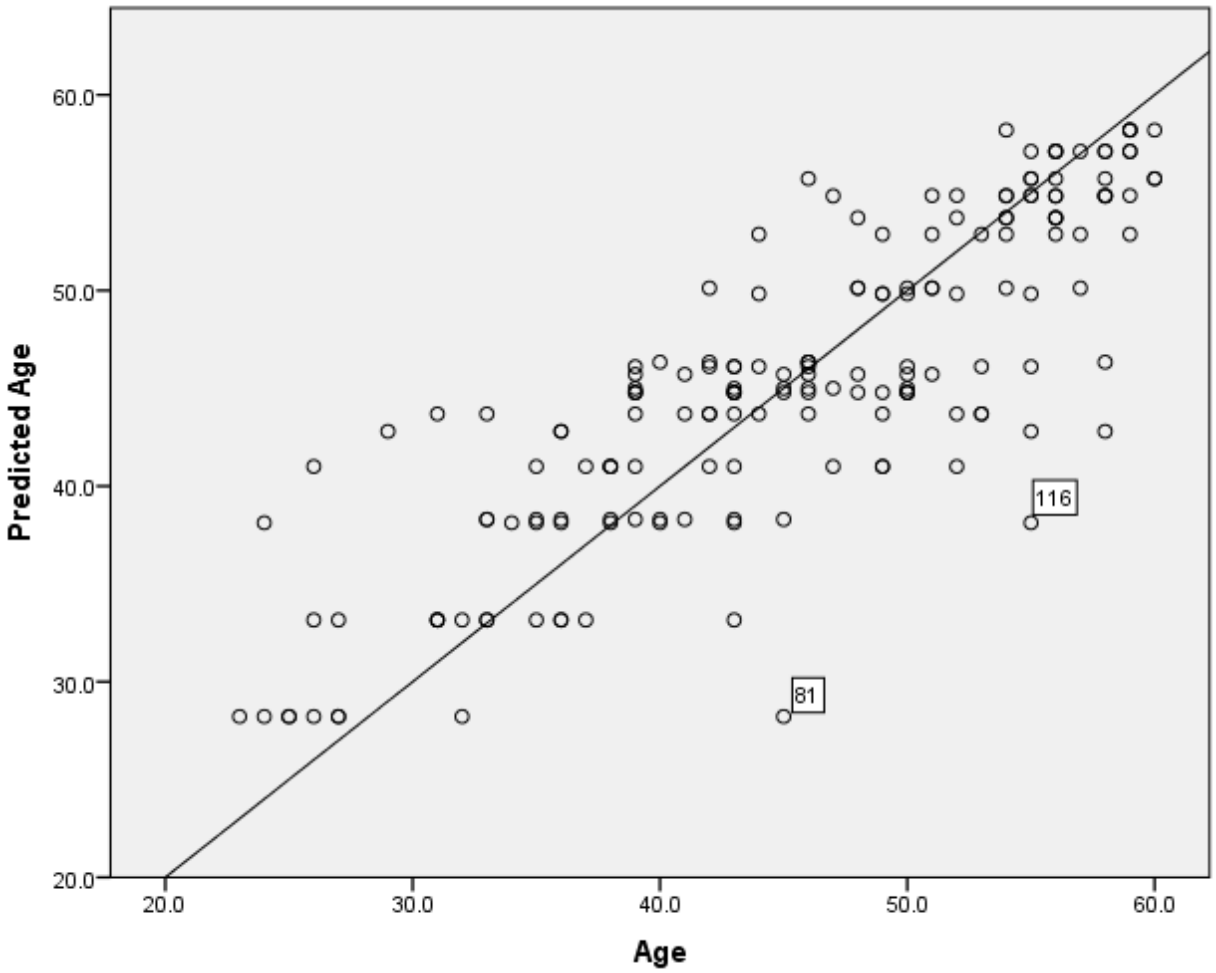


Figure 5.3: Predicted ages of CHAID 10/5 with all observations plotted against actual age. The two largest residuals are highlighted.

The second decision tree shows slightly better goodness-of-fit measures with a mean absolute error of 3.6 years and explaining about 74.15% of the variance in age. This is simply an artifact of the reduction of the variance in age by the removal of two outlying observations.

In a study such as this with a high number of variables, it is highly likely that at least one (if not more than one) significant variable would be found in an analysis. This is shown by how far the CHAID 10/5 decision tree splits. With very small terminal nodes, the tree basically split until all the nodes had reached minimum parent node sizes. Under these circumstances, a

decision tree that is allowed to grow will grow extensively. The limiting growth factor becomes the sample size and the minimum parent and child node size parameters. Therefore, good values for the parent and child node minimum size parameter have to be selected in order to get a good decision tree. It needs to be large enough to allow for valuable interpretations and conclusions, but not so large that it becomes too data specific and too hard to draw any valuable conclusions. In this study, the parent and child minimum node size settings are utilized to prevent the decision tree from growing excessively large. Generally, the terminal node size will lie between the parent and child minimum node values. It is possible that a terminal node will be larger than the minimum parent node size. This means one of two things: either all the observations in that node are so homogenous that no variables could be found to split them further, or that any subset of the observations distinguishable from the rest is smaller than the minimum child node size value.

Neither the CHAID 100/50 nor the CHAID 10/5 decision trees seemed very useful to draw conclusions about the relationship between age indicators and actual age. The parent and child minimum node size parameters were changed to find values that would create decision trees with nodes large enough not to be affected by outliers, and enough branches to allow for some interesting interpretation. Several sets of values for the parent and child minimum node size were tried. The decision trees that resulted from using the values 30 and 10 appeared to represent the best compromise between fitting the data too specifically and still having the tree be large enough to show some variable relationships between the age indicators and between indicators and age. A selection of the other CHAID trees can be found in the appendix (Figures C.2, C.3, and C.4). These trees are very similar to the chosen middle-of-the-road decision tree CHAID 30/10 presented in Figure 5.4. For instance, CHAID 30/5 (Figure C.3) is an exact match

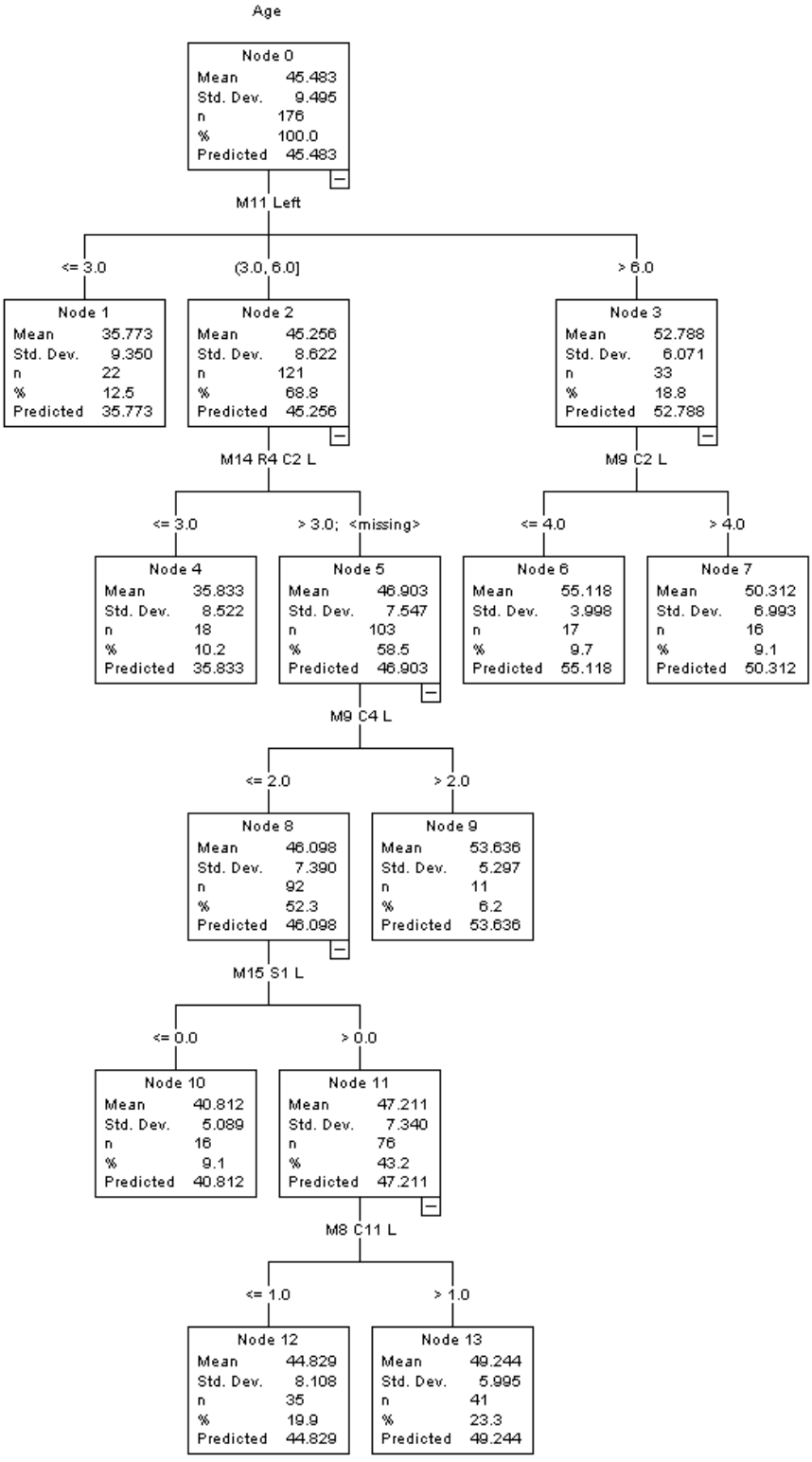


Figure 5.4: CHAID 30/10 with all variables and observations

to CHAID 30/10 and simply adds three splits onto the bottom of CHAID 30/10. If you split node 1 of CHAID 30/10 using dorsal pubis scoring from Gilbert and McKern (1973) the resulting tree becomes CHAID 20/10 (Figure C.2). Similarly, if you drop the split under node 3 in CHAID 30/10 based on measurements of the auricular surface texture (M9 C2 L) (Buckberry and Chamberlain 2002), the result is identical to CHAID 40/10 (Figure C.4).

For all these decision trees, the central part of the tree remains the same, represented by CHAID 30/10 (Figure 5.4). It uses six variables to create eight age groups. This model evaluates with a mean absolute error of prediction of 5.57 years and accounts for approximately 47% of the variance in age. Probably most interesting is that out of 94 variables used, none of the 39 pubic symphysis age indicator variables were chosen by the algorithm as being closely related to the data. Four of the variables chosen pertain to the auricular surface (Lovejoy phases (1985b), Buckberry and Chamberlain components of macroporosity and surface texture (2002), and inferior surface texture (Boldsen et al. 2002)). The other two variables are the midlambdoidal suture closure (Meindl and Lovejoy 1985) and the sternal rib end pit shape described by İşcan et al. (1984b). The expectation would be to see at least one variable for each of the major age indicator sites, with additional variables used to fine tune the model. In this decision tree the most important variables are the auricular surface phase (Lovejoy et al. 1985b), followed by the pit shape of the ribs (İşcan et al. 1984b), and then the amount of macroporosity observable on the auricular surface (Buckberry and Chamberlain 2002). The other variables that are present in this tree are less influential because they pertain to smaller groups of observations. While the shape of the rib pit (M14 R4 C2 L) influences 121 observations, the surface texture of the auricular surface (M9 C2 L) (Buckberry and Chamberlain, 2002) only pertains to 33 observations.

Figure 5.5 shows the predicated ages against the actual ages for CHAID 30/10 and highlights the three observations with the largest residual values. The difference between actual and predicted age is much larger for these three observations when compared to the others. Excluding the highest residuals most of the ages were estimated to within 15 years. In order to test the influence of the observations with the large residuals they needed to be removed to see if the model would change significantly.

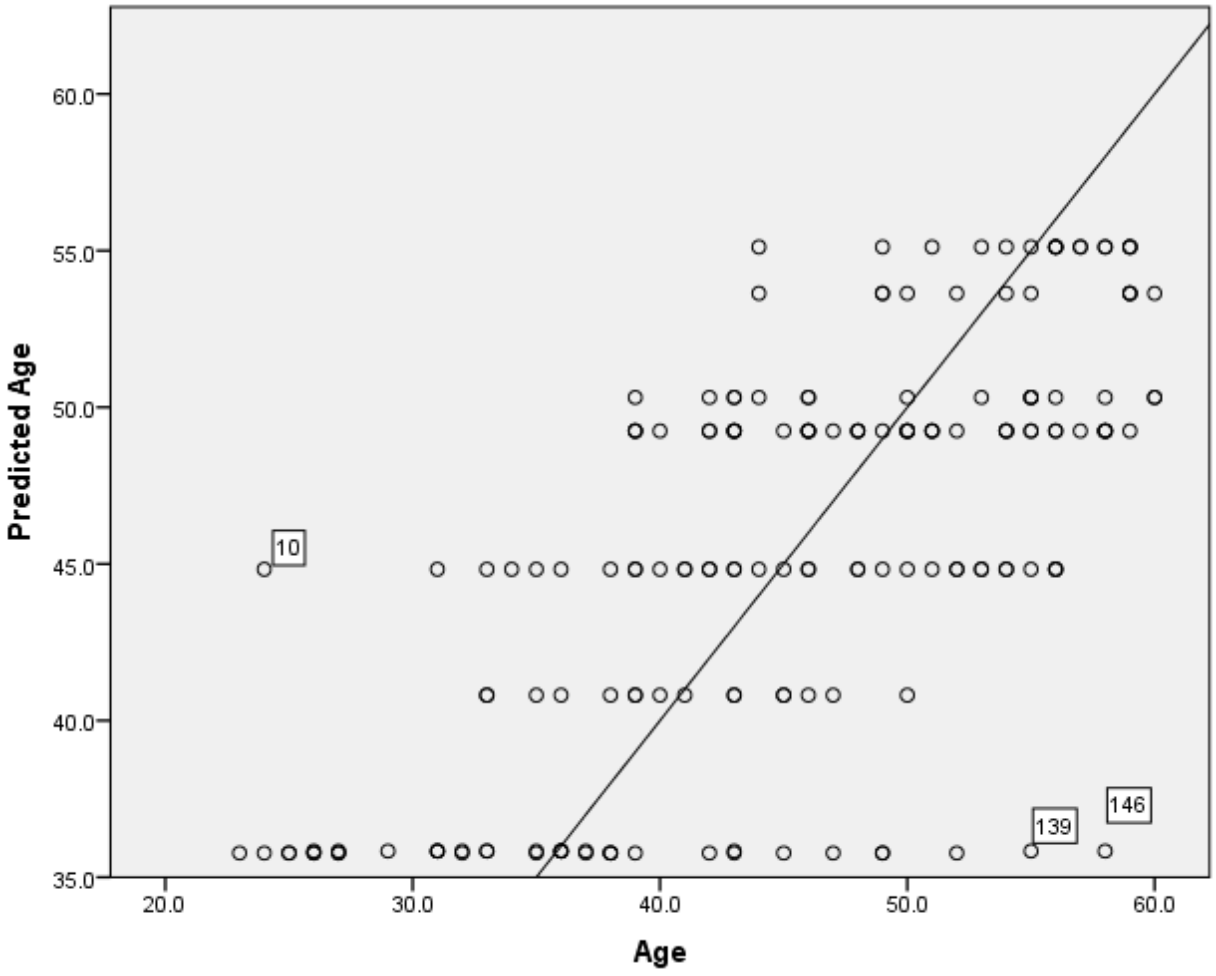


Figure 5.5: Predicted ages of CHAID 30/10 with all observations plotted against actual ages

In the case of CHAID 30/10, when the most extreme observation (observation 146) is removed and the analysis is rerun, the resulting decision tree has the same structure with the same variables, still not selecting any pubic symphysis indicator variables. The new model has an average absolute error of 5.46 years which is basically the same as the original model. The proportion of the variance in age accounted for in the model is 49.84%, which is also only slightly higher than the original.

Figure 5.6 shows the plot of predicted ages against actual ages for the decision tree model without observation 146. The two observations with the largest residual values are highlighted in the graph. Of these two observations, observation 10 has a slightly larger absolute error than observation 139. Removing observation 10 in addition to observation 146 changes the resulting CHAID 30/10 decision tree slightly. Shown in Figure 5.7, the tree has the same structure as the previous two, but changes a little bit toward the bottom. Here, the new tree adds two new splits using the traverse palatine suture closure as scored by definitions of Meindl and Lovejoy (1985), and the development of the ventral bevel of the pubic symphysis as redefined by this author based on definitions laid out by Chen et al. (2008). However, the main core of the tree stays the same, with the phase of the auricular surface and shape of the rib pit still being the two first and most important variables in this decision tree. The two new splits at the bottom of the tree increase the performance of the model noticeably. The new model now accounts for 53.68% of the variance in age and the observations show a mean absolute error of 5.14 years. The additional variables chosen in this decision tree after the removal of observation 10 suggest that this observation exerted some minor influence at the bottom of the previous decision tree.

The residuals of this model still show observation 139 as being unusually large compared to the other residuals. Removing observation 139 in addition to observations 10 and 146 has a

profound effect on the resulting decision tree. Figure 5.8 shows the CHAID 30/10 tree resulting when observations 10, 146, and 139 are not used to construct the model. The resulting tree looks nothing like any of the previous CHAID 30/10 decision trees.

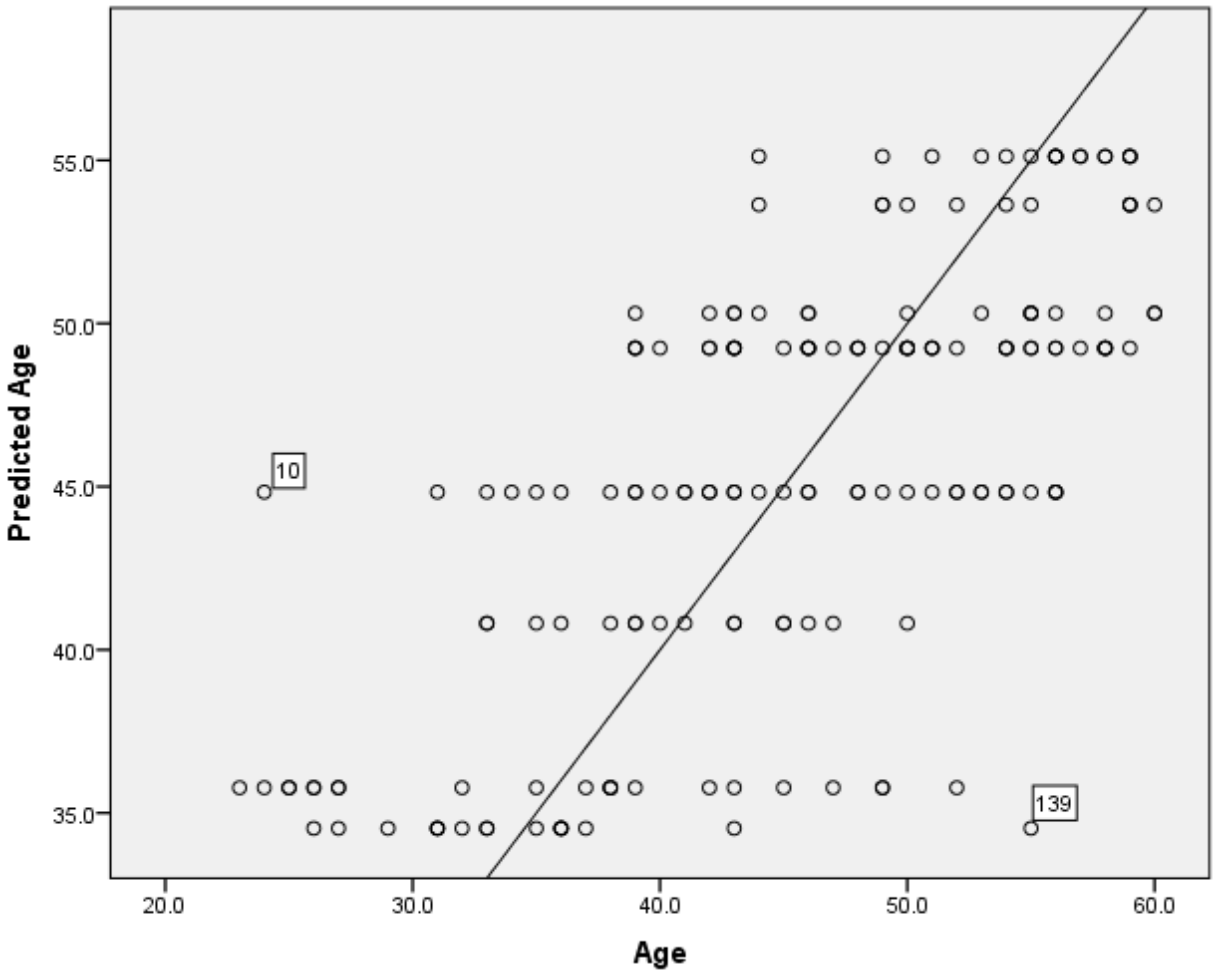


Figure 5.6: Predicted ages of CHAID 30/10 without observation 146 plotted against actual ages

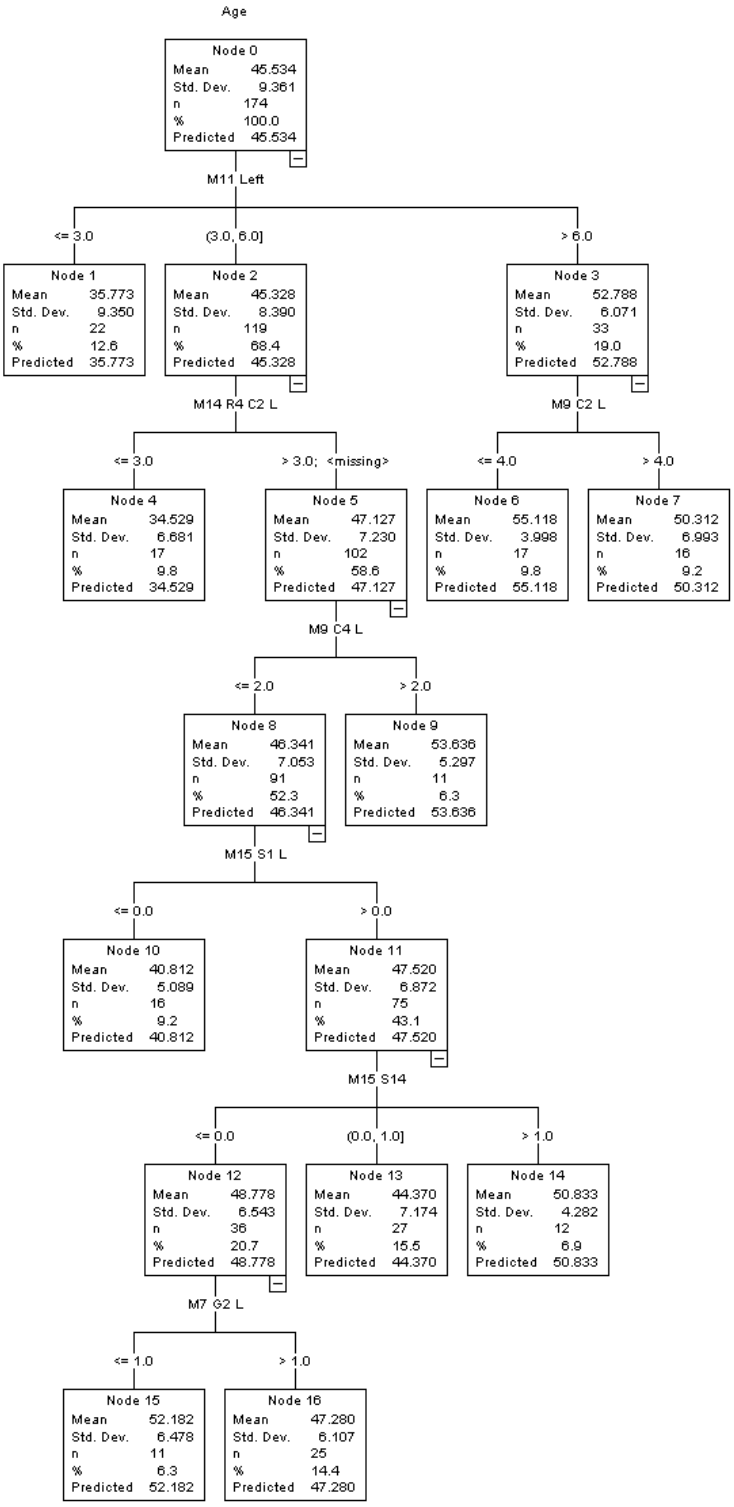


Figure 5.7: CHAID 30/10 without observations 10 and 146

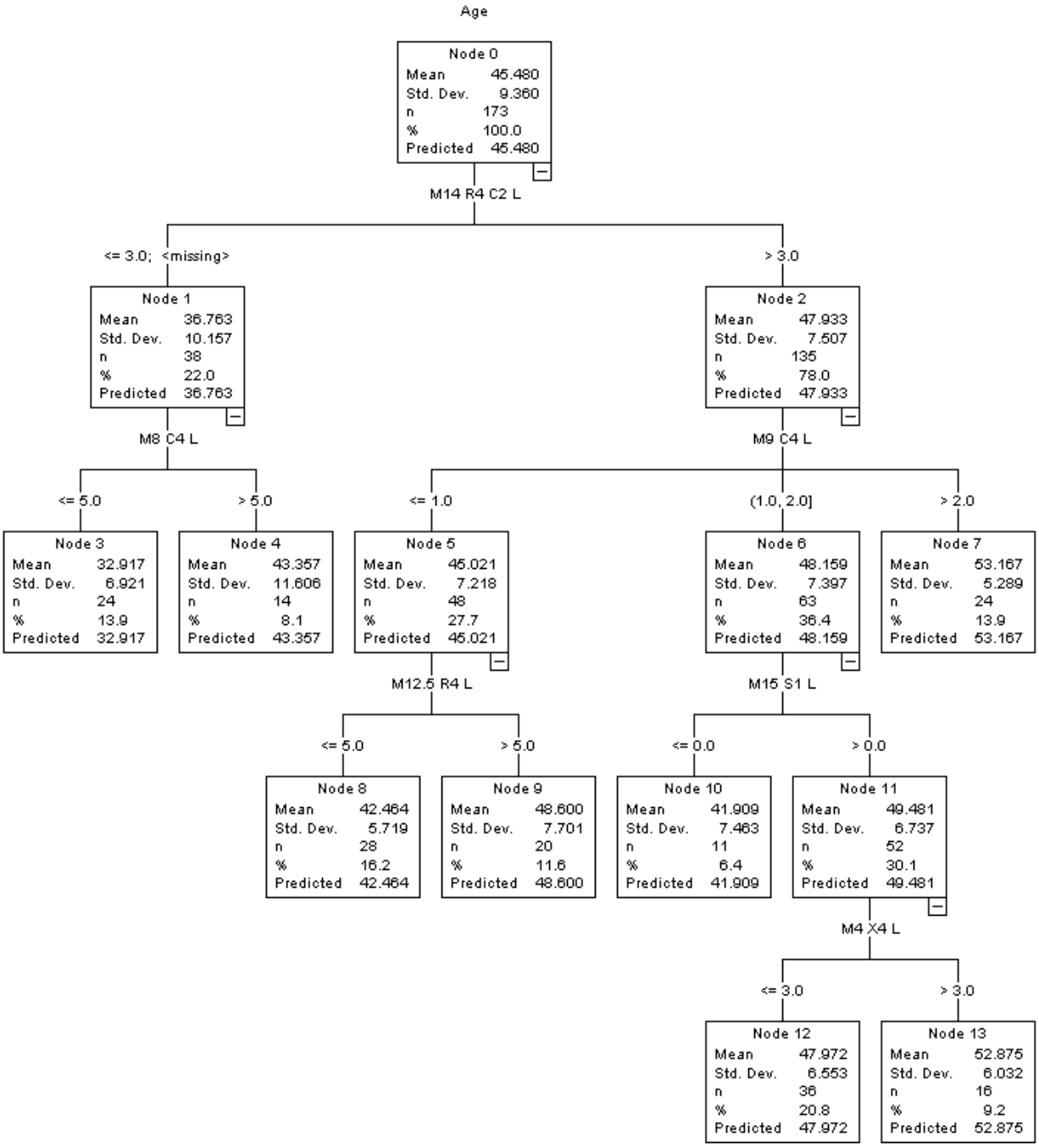


Figure 5.8: CHAID 30/10 without observations 10, 139, and 146

The new decision tree shows a different pattern of age indicator variables. Instead of the auricular surface phase dictating the initial split in the tree, the pit shape of the rib now defines the first split. Since this split is based on a different variable, it follows that the child nodes created from this split would be different from before as well. In decision trees, the growing algorithm chooses variables and splitting rules based on the observations in the node under scrutiny. In this tree, the initial split is different and creates different child nodes. Therefore, it is easy to understand how nothing under the first difference can look the same either. Once one change in decision trees occurs, either by a different variable or a different splitting value, everything after that change will also change.

The CHAID 30/10 decision tree without observations 10, 139, and 146 performs about the same as the original CHAID 30/10, but slightly worse than the subsequent versions. This model accounts for 46.78% of the variance in age and predicts ages with a mean absolute error of 5.55 years. The tree uses six variables to create eight age groups. The primary variables of importance in this model are the pit shape of the rib (İşcan et al. 1984b), the amount of auricular surface macroporosity (Buckberry and Chamberlain 2002), and suture closure at midlambdoid (Meindl and Lovejoy 1985). The variables of lesser influence are the sex-specific phase of rib end morphology (İşcan et al. 1984b; İşcan et al. 1985), development of the dorsal aspect of the pubic symphysis (Hanihara and Suzuki 1978), and the development of the ventral aspect of the pubic symphysis (Boldsen et al 2002). Unlike the other CHAID 30/10 decision trees which did not utilize any information from the pubic symphysis, this tree uses two variables that pertain to age-related changes of the pubic symphysis. It also uses two variables based on the ribs, one on sutures, and one on the auricular surface.

Figures 5.9 through 5.13 show various graphs of the residuals of the last CHAID 30/10 decision tree missing observations 10, 139, and 146. These graphs show how the residuals relate to the predicted ages and the actual ages. Several patterns that reappear throughout all of the decision trees in this study are apparent in these graphs. In Figure 5.9, the residuals are plotted against actual age. Ideally, the pattern should be an even scattering on the horizontal axis. However, the pattern this graph shows is that the method consistently overages young individuals and underages old individuals. Positive residuals occur when the estimated age is smaller than the actual age, therefore estimating an individual to be younger than they actually are. Alternatively, negative residuals indicate individuals estimated to be older than they actually are. In Figure 5.10, age is plotted against the absolute value of the residuals, which only looks at the size of the error and ignores the direction. This graph shows an outward flaring of the scatter giving it the look of a sideways "V." This pattern indicates that larger residuals are associated with individuals at either end of the age range of the data (i.e., the older and younger individuals). Figure 5.11 shows the residuals graphed against the predicted values. In this graph, there is a nice even scattering around the 0 mark meaning that the age estimates are as likely to underage as they are to overage no matter the predicted age. Figure 5.12 shows the absolute residuals graphed against the predicted values. In this graph, almost all of the errors are less than 15 years, with most errors being even less than approximately 12 years. The graph shows the highest residuals to be equally common for all predicted age values so the predictions are equally accurate, or inaccurate, for all ages. If there was a pattern evident of tapering toward the top, the graph would indicate that younger estimated ages have a higher uncertainty associated with them than older age estimations. Lastly, Figure 5.13 graphs the actual ages against the predicted ages. This shows a general upward trend, indicating that older individuals are estimated to be older.

This is a good trend; however, in Figure 5.13 the trend is not very precise. The spread of the actual ages for any particular predicted age value is very wide, meaning that an individual predicted to be 43 years old could actually be between 25 and 58 years. If the decision tree were very good at age estimation, we would expect to see a very narrow horizontal scattering of data points in this graph.

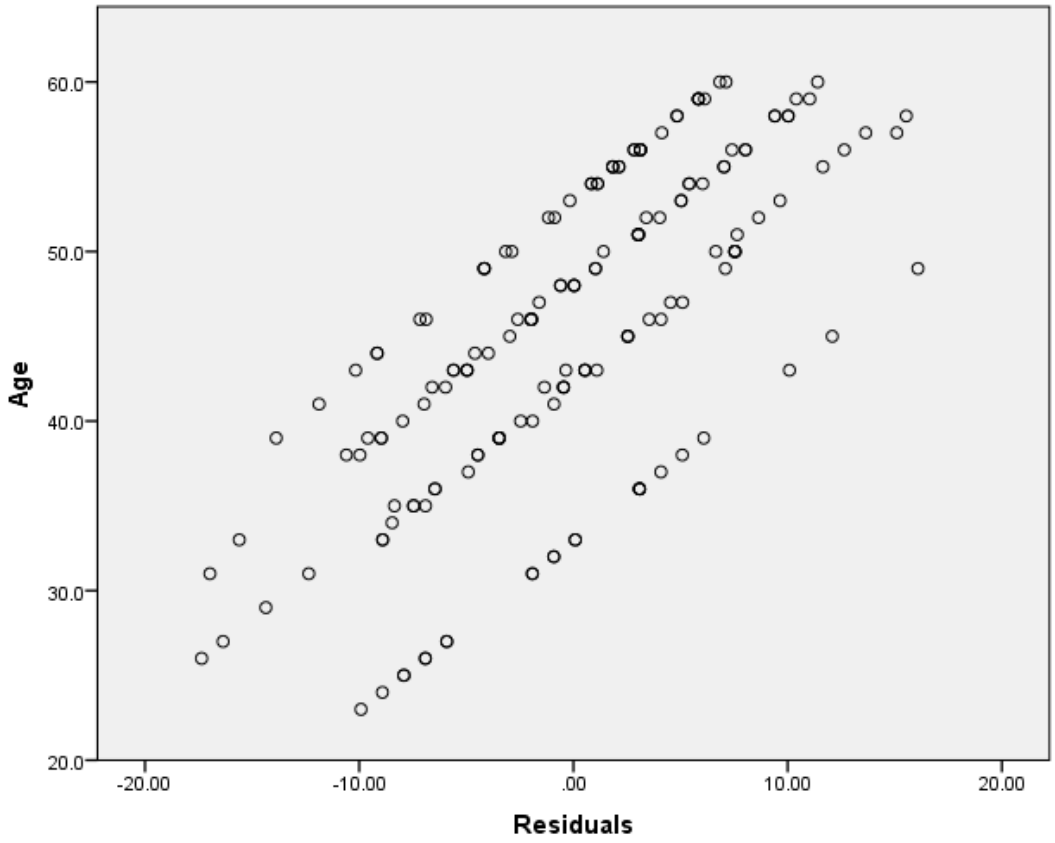


Figure 5.9: Residuals of CHAID 30/10 without observations 10, 139, and 146 plotted against actual age

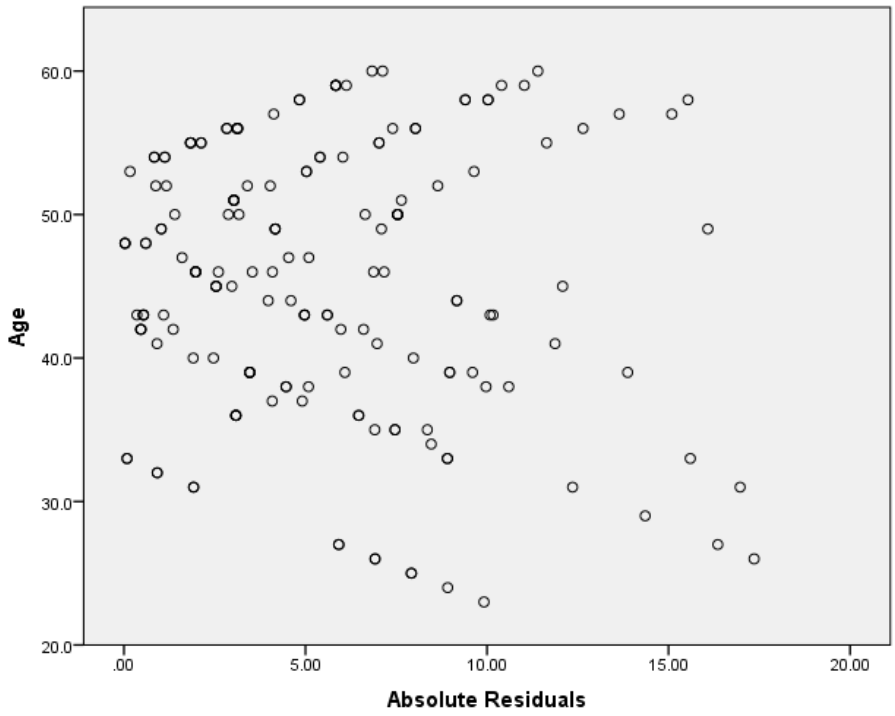


Figure 5.10: Absolute residuals of CHAID 30/10 without observations 10, 139, and 146 plotted against actual age

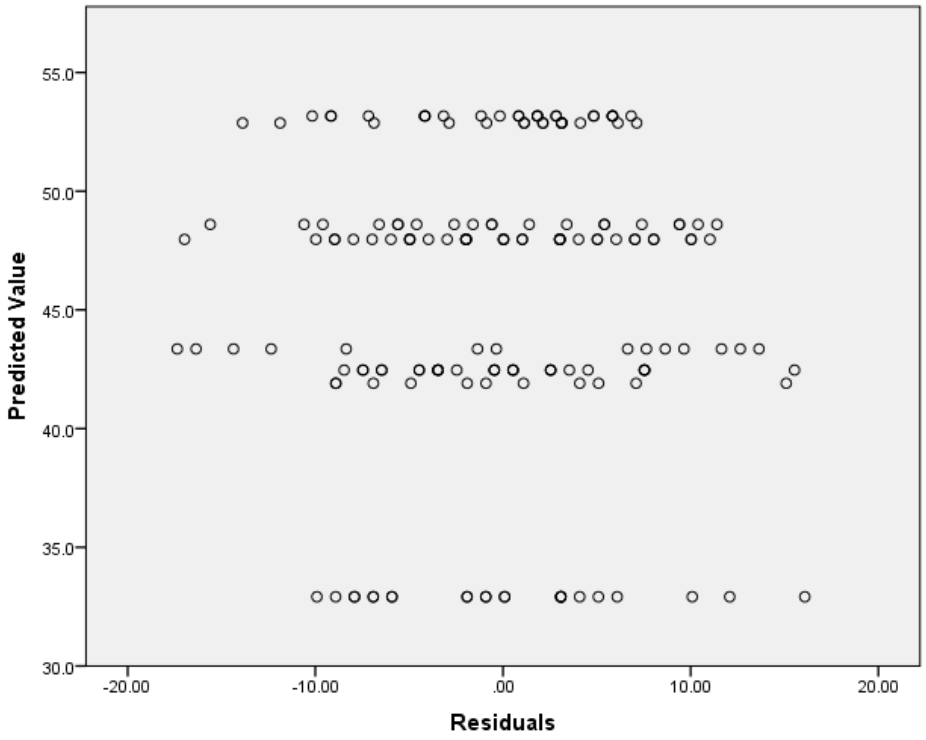


Figure 5.11: Residuals of CHAID 30/10 without observations 10, 139, and 146 plotted against predicted age

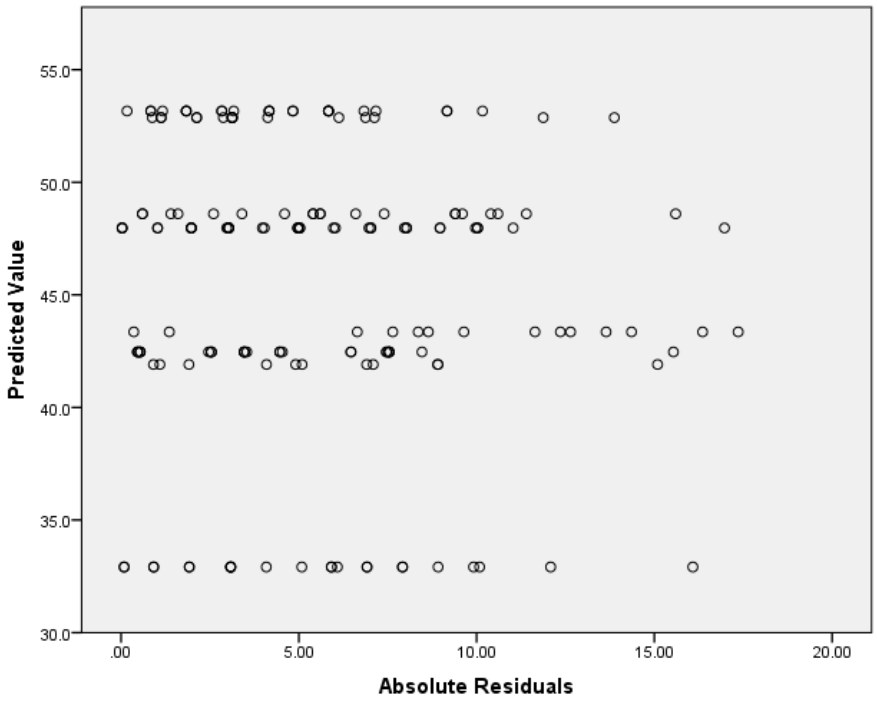


Figure 5.12: Absolute residuals of CHAID 30/10 without observations 10, 139, and 146 plotted against predicted age

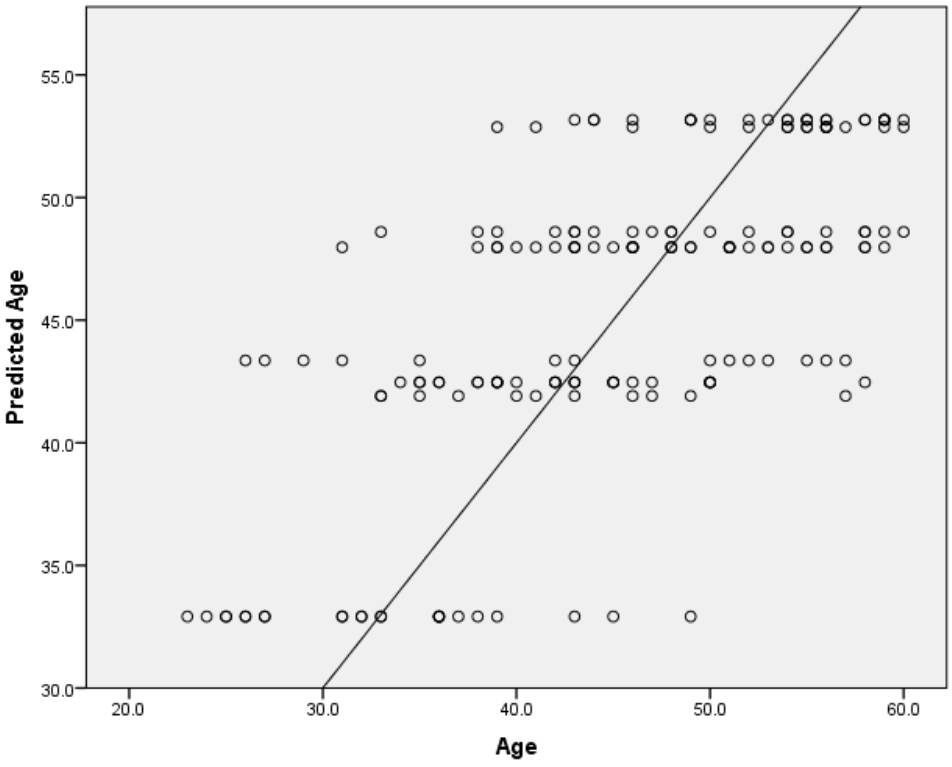


Figure 5.13: Predicted ages of CHAID 30/10 without observations 10, 139, and 146 plotted against actual ages

Table 5.1: Performance Summary Values for various CHAID decision trees.

Model	Mean Absolute Error	% Variance
CHAID 100/50 all vars, all obs	7.004	19.31%
CHAID 10/5 all vars, all obs	3.8105	70.25%
CHAID 10/5 all vars, w/o obs 81, and 116	3.5967	74.15%
CHAID 20/10 all vars, all obs	5.3285	50.98%
CHAID 30/5 all vars, all obs	5.2127	52.28%
CHAID 40/10 all vars, all obs	5.6747	45.85%
CHAID 30/10 all vars, all obs	5.5672	47.05%
CHAID 30/10 all vars, w/o obs 146	5.4554	49.84%
CHAID 30/10 all vars, w/o obs 10, and 146	5.1409	53.68%
CHAID 30/10 all vars, w/o obs 10, 139 and 146	5.5505	46.78%

Table 5.1 shows the various CHAID trees utilizing all variables with their associated mean absolute errors and the proportion of the variance in age accounted for by the model. The best age estimation performance, excluding the huge CHAID 10/5 tree, goes to the CHAID 30/10 tree excluding observations 10 and 146. It also shows the highest explained variance. Second place goes to CHAID 30/5. However, most are very similar. Large differences in the performance are predominantly caused by the size of the decision trees, with the large extensive trees outperforming the smaller ones.

Classification and Regression Tree (CRT) Decision Trees

CRT decision trees were also constructed on the data set. CRT trees are constructed by finding splits in the data that create highly homogenous groups. Therefore, CRT trees are more influenced by the structure of the data, rather than the variables. Where CHAID trees select their trees based on chi-squared correlations between the age indicators and actual age, the CRT algorithm is driven more directly by the ability to find large homogenous subgroups of data. CRT decision trees always make binary splits, so that every parent node will give rise to exactly

two child nodes. Some nodes in the CHAID trees of the previous section had triplets. The SPSS version of the CRT decision tree construction also has the option of using surrogates and pruning the decision tree. Surrogates refers to the treatment of missing data when building the CRT tree. When using surrogates, missing data is estimated from variables that are similar to the variable with the missing values. Pruning refers to eliminating some insignificant splits in the decision tree after it is grown to its full size. Neither option was utilized in this study for reasons previously discussed in the methods chapter.

The analysis of CRT decision trees started the same way the CHAID analysis started, constructing a CRT tree using the defaults in the SPSS analysis. The default tree was CRT 100/50. The results were the same with and without the use of surrogates. The resulting tree was very simple—far too simple as it turned out. It had only one node containing all the observations and therefore could not really be considered a decision tree. In this model, no decisions had to be made and all observations were estimated to be the same age. For subsequent analyses, the option to use surrogates was turned off in order to limit the obscuring effect surrogates might have on the interpretation of the resulting CRT decision trees.

As with CHAID trees, CRT 10/5 was then grown to see the extent of maximum tree growth allowed by relaxing the constraints on minimum parent and child node size. The resulting decision tree is shown in Figure 5.14. CRT 10/5 is large and complex, much like CHAID 10/5, and uses 14 age indicator variables to create 15 age groups. With an average absolute error of 4.45 years, the decision tree accounts for 67.87% of the variance in age.

The variables selected in this decision tree are of particular interest. Out of the 14 variables chosen, only a single one (M4 X4 L) pertains to the pubic symphysis and is of very minor importance to the model. Of the other 13 variables, two pertain to the ribs, five to cranial

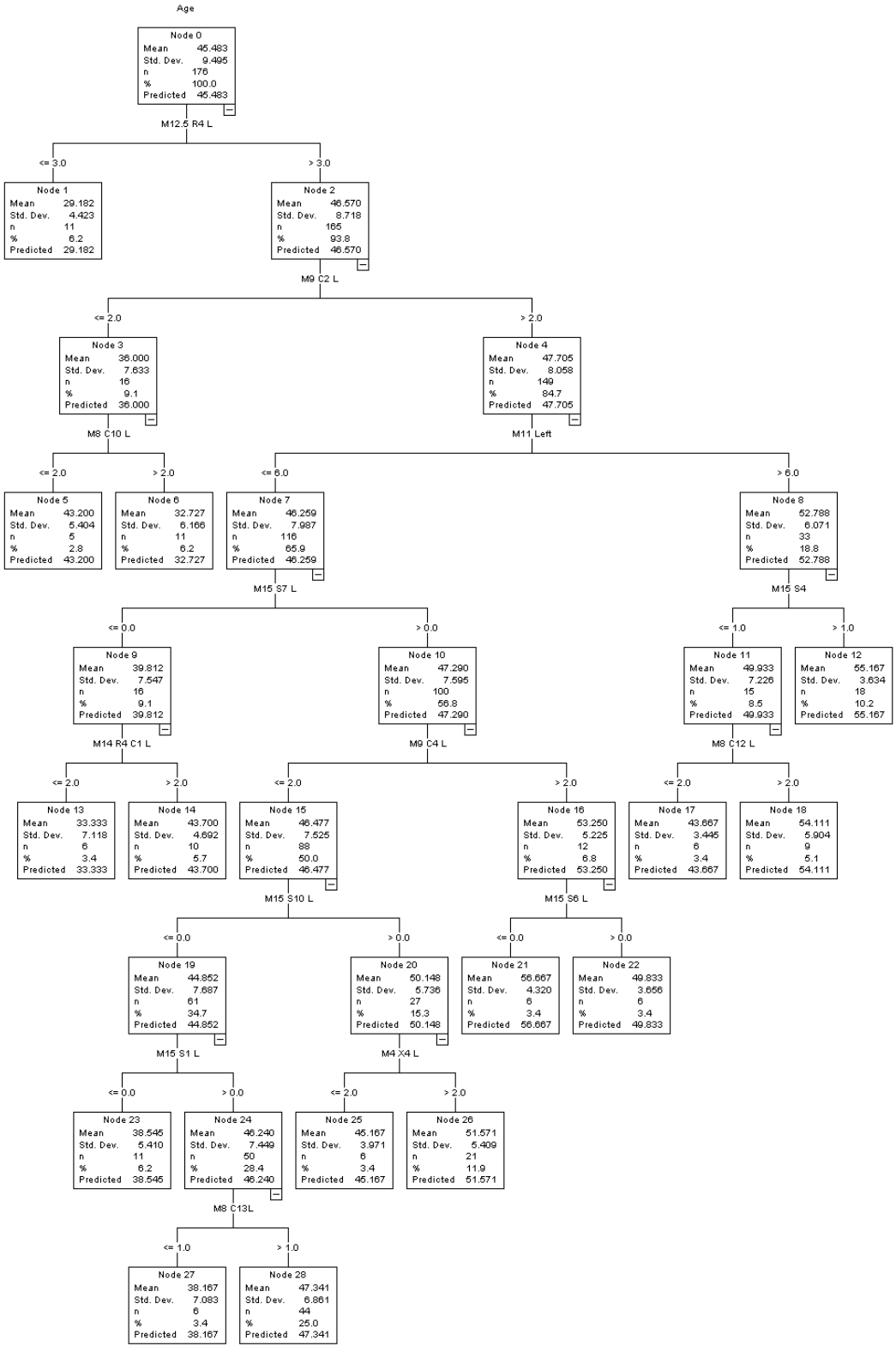


Figure 5.14: CRT 10/5 decision tree with all variables and all observations

sutures, and six to the auricular surface.

The variables used in this decision tree are the sex-specific phases of the sternal rib end (İşcan et al. 1984a; İşcan et al. 1985), pit depth of the sternal rib end (İşcan et al. 1984b), phase of the auricular surface (Lovejoy et al. 1985b), degree of macroporosity and surface texture of the auricular surface (Buckberry and Chamberlain 2002), morphology on the inferior auricular surface, superior posterior iliac exostoses and inferior posterior iliac exostoses (Boldsen et al. 2002), suture closure at midlambdoidal, anterior sagittal, midcoronal, pterion, and superior sphenotemporal (Meindl and Lovejoy 1985), and the development of the dorsal aspect of the pubic symphysis (Hanihara and Suzuki 1978).

Unlike the CHAID decision trees which predominantly selected the auricular phase as the most important variable, in the CRT tree this variable is pushed into third place and is supplanted by the phase of the ribs and the texture of the auricular surface. Auricular macroporosity and suture closure at pterion and superior sphenotemporal are also variables at major splits in the decision tree. The other eight variables are of lesser importance controlling the fates of fewer observations.

Figure 5.15 shows the predicted ages against actual age. No extreme residuals were noted. While observation 15 was the largest, it did not lie outside the range of normal expectation and was not flagged as an extreme residual. To test the impact of this observation on the model, it was removed nonetheless and the model was rerun.

The resulting decision tree is slightly smaller than CRT 10/5 with all observations. Reproduced in Figure C.5 in the appendix, it only utilizes nine variables to group the observations into 10 skeletal age cohorts. For this smaller model, the mean absolute error was 5.36 years with the model accounting to 50.01% of the variance in age. The tree is very

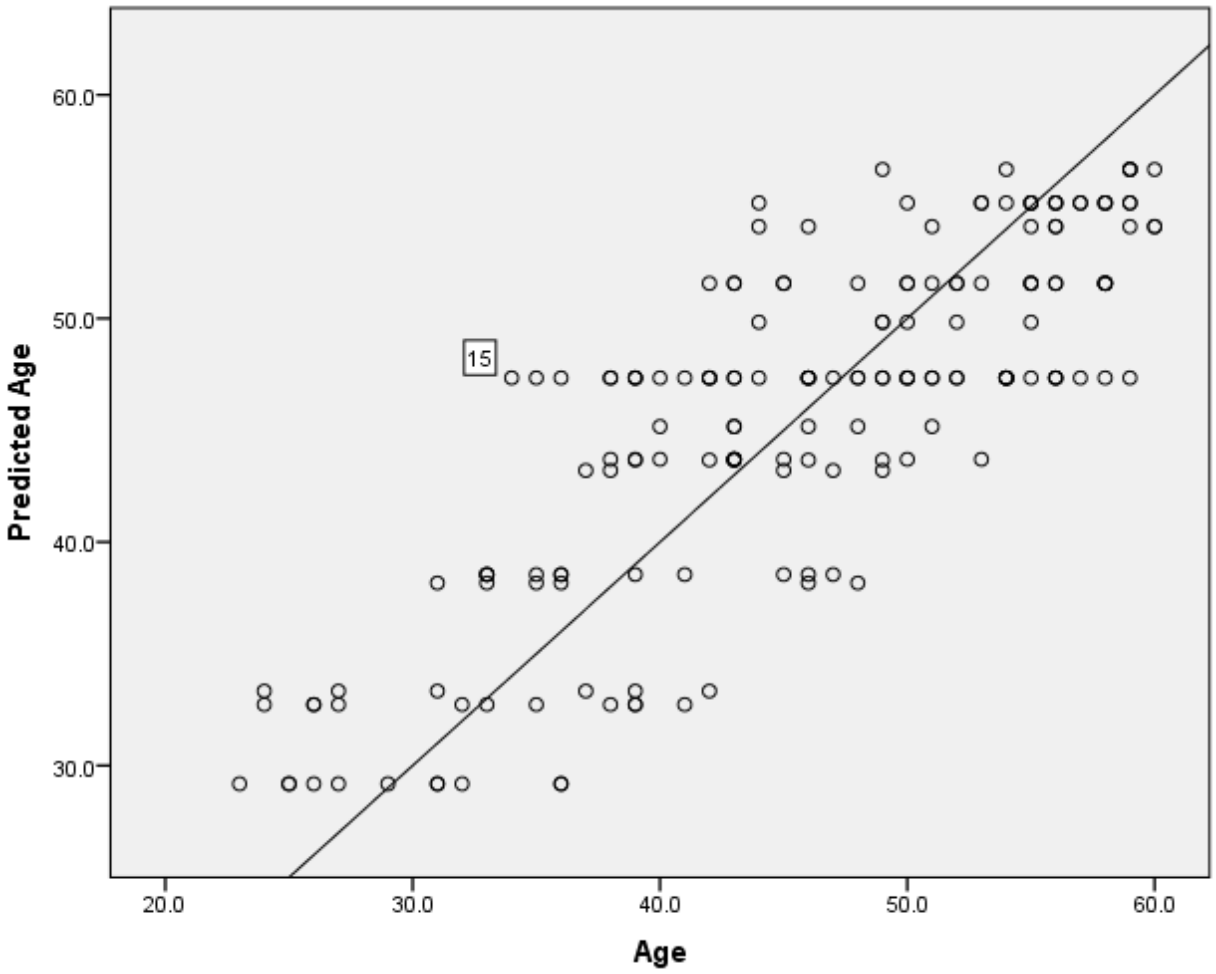


Figure 5.15: Predicted ages of CRT 10/5 plotted against actual ages

symmetrical and probably the prettiest in the whole study. The variables in this decision tree are similarly distributed across the spectrum of age indicator sites as the complete CRT 10/5 decision tree. The pubic symphysis is only represented by a single variable and appears fairly far down the tree. The auricular surface is represented by four variables and the ribs and cranial sutures each have two variables. Specifically, these variables are pit depth and pit shape of the sternal rib ends (İşcan et al. 1984b), suture closure at midcoronal and sphenofrontal (Meindl and Lovejoy 1985), surface texture of the pubic symphysis (Chen et al. 2008), degree of microporosity on the auricular surface (Buckberry and Chamberlain 2002), morphology near the

apex and the inferior portion of the auricular surface, and the inferior posterior iliac exostoses (Boldsen et al. 2002). The most important variables to the model are the rib pit shape, auricular microporosity, and the morphology of the auricular surface. This tree exhibits a much different structure than the CRT 10/5 constructed on the complete data set.

As with the CHAID 10/5 decision tree, both CRT 10/5 decision trees seemed to be subject to overfitting as they exhibit very small terminal nodes. In order to adjust the size of the decision tree the minimum parent and child node sizes were expanded to find a tree with good interpretability. For this data set, the CRT growing algorithm refused to split if the minimum child node size was set to anything greater than 12. That means that even CRT 20/12 was the same as CRT 100/50. This suggests that the data does not have large homogenous groups, which is attributable to the large number of variables used in this data set. The more variables are used, the less likely it becomes for two observations to have matching values across all the variables. Only when the minimum child size was lowered to 11 did any CRT decision trees grow. However, even then the decision trees were only "subtrees" of CRT 10/5 with some of the lower, smaller splits removed. The extent to which the CRT tree would grow was determined by the minimum parent node size. CRT 50/10 is reproduced in appendix Figure C.6. When this decision tree is compared to CRT 10/5 in Figure 5.14, the similarity becomes apparent very quickly. CRT 50/10 is simply the central "root" of CRT 10/5. As such all the variables of CRT 50/10 are a subset of the variables of CRT 10/5. None of the variables in the CRT 50/10 decision tree are associated with the pubic symphysis. It only uses seven variables to create eight age groups, which is about half the amounts of the maximum CRT 10/5 decision tree with its 14 variables and 15 groups. As it is only a section of CRT 10/5, the reduced decision tree CRT 50/10 has

worse goodness-of-fit measures with an average absolute error of 5.34 years and accounts for 54.18% of the variance in age.

However, CRT 50/10 is very linear and a slightly more complex tree was chosen as the model of the CRT decision tree evaluation. Tree CRT 30/10, shown in Figure 5.16, has one more split than CRT 50/10. It uses eight variables to delineate nine groupings of observations. With a mean absolute error of 5.19 years and accounting for 55.59% of the variance in age, it represents only a slight improvement over the simpler CRT 50/10 decision tree. The variables that remain in CRT 30/10 as important are the sex-specific phases of the rib ends (İşcan et al. 1984a; İşcan et al. 1985), phase of the auricular surface (Lovejoy et al. 1985b), degree of macroporosity and surface texture of the auricular surface (Buckberry and Chamberlain 2002), and the state of suture closure at midlambdoidal, anterior sagittal, pterion, and superior sphenotemporal (Meindl and Lovejoy 1985). Looking at the graph of predicted ages against actual ages in Figure 5.17, no obvious outlying observations are visible. In terms of magnitude, the errors generally do not exceed 15 years at the most, and the most extreme errors are associated with observations 25 and 10.

Given that the CRT decision trees constructed up to this point are all related to one another (as they are portions of the CRT 10/5 decision tree), it appears that the set of all 176 observations will not change the basic structure of the tree or select different variables. It was also shown that removing some observations and growing the CRT 10/5 decision tree again resulted in some structural changes to the tree, including a different set of variables (compare Figures 5.14 and C.5). In order to find a larger set of important variables as selected by the CRT algorithm, the most extreme observation (observation 10) of CRT 30/10 was removed and a new

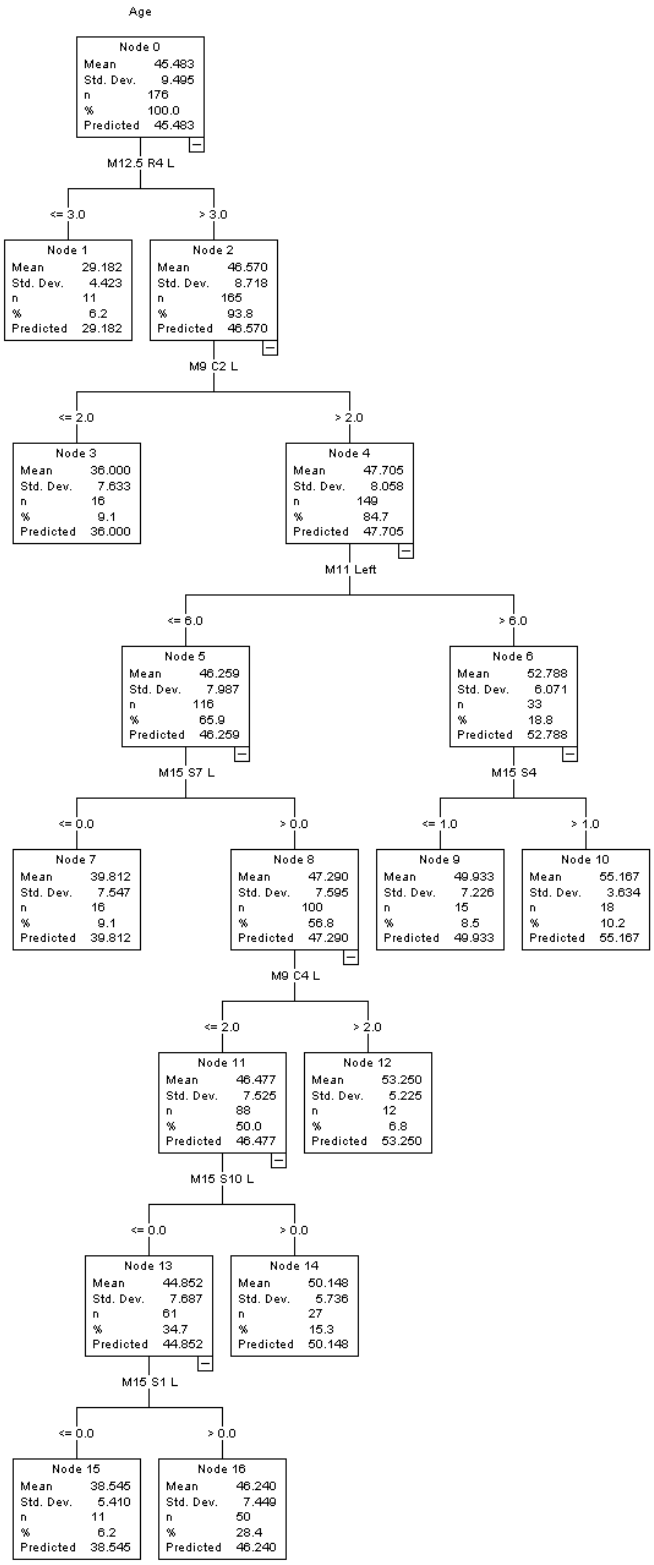


Figure 5.16: CRT 30/10 with all variables and all observations

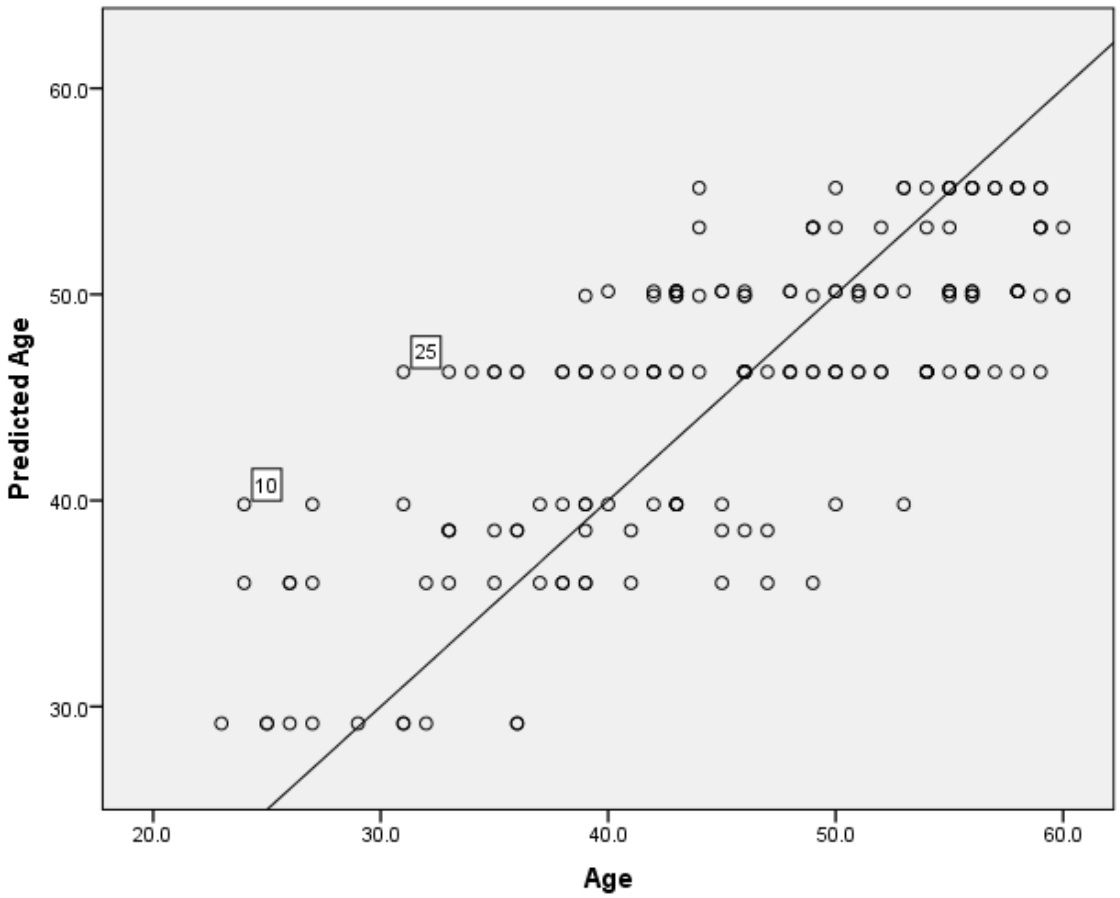


Figure 5.17: Predicted ages of CRT 30/10 with all variables and observations plotted against actual ages

CRT 30/10 tree was grown. This new decision tree, shown in Figure 5.18, is slightly simpler than the CRT 30/10 with all observations using only six variables to create seven groups.

The tree excluding observation 10 does not fit the data as well as CRT 30/10, only accounting for 40.43% of the variance in age and producing a mean absolute error of 5.67 years. The variable set of this decision tree is very different than that found in CRT 30/10, but bears a resemblance to the CRT 10/5 decision tree that was constructed without observation 15 (see Figure C.5). The tree used variables pertaining to each of the four age determination sites used in this study—even the pubic symphysis.

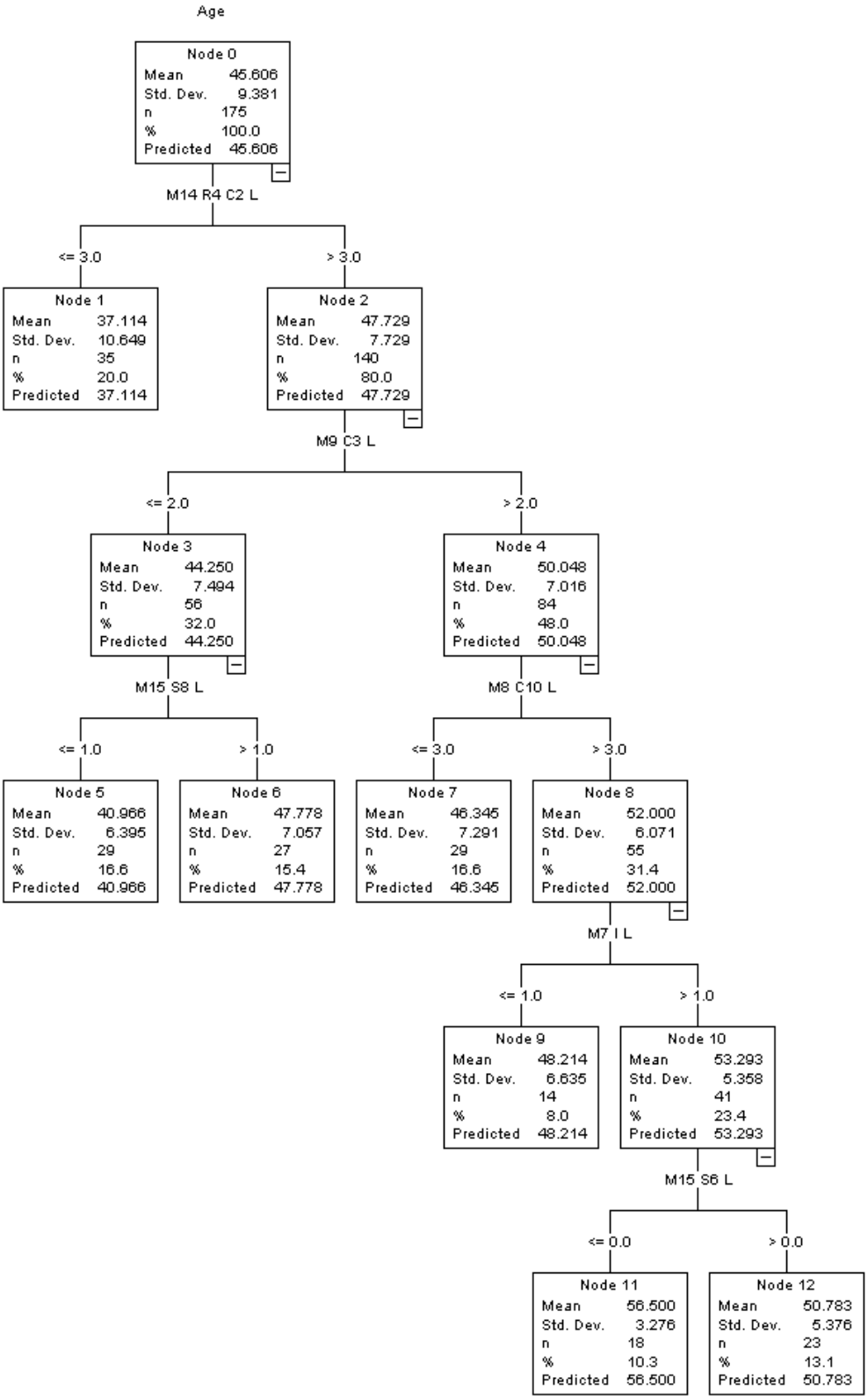


Figure 5.18: CRT 30/10 with all variables and excluding observation 10

The variables in the CRT 30/10 decision tree without observation 10 are the pit shape of the rib ends (İşcan et al. 1984b), the degree of microporosity on the auricular surface (Buckberry and Chamberlai, 2002), the morphology of the inferior aspect of the auricular surface (Boldsen et al. 2002), the surface texture of the pubic symphysis (Chen et al. 2008), and the degree of suture closure at midcoronal and sphenofrontal (Meindl and Lovejoy 1985). The top three variables are the shape of the rib pit, the amount of auricular microporosity, and morphology of the inferior auricular surface.

Further removal of the other noted observation in Figure 5.17 from CRT 30/10 (observation 25) in addition to observation 10 produced the same decision tree as when only observation 10 is removed. Neither of these decision trees shows any further extreme outlying observations.

Table 5.2 shows a summary of the performance of the various CRT decision trees run using all 94 variables in the data set. CRT 30/10 with all observations has the best performance numbers with the exception of the complete CRT 10/5 decision tree which is deemed too data set specific to provide reliable interpretative information.

Table 5.2: Performance Summary Values for various CRT decision trees.

Model	Mean Absolute Error	% Variance
CRT 10/5 all vars, all obs	4.4538	67.87%
CRT 10/5 all vars, w/o obs 15	5.3610	55.01%
CRT 50/10, all vars, all obs	5.3425	54.18%
CRT 30/10, all vars, all obs	5.1850	55.59%
CRT 30/10, all vars, w/o obs 10	5.6711	40.43%
CRT 30/10, all vars, w/o obs 10, 25	5.6538	40.26%

Decision Trees using Variable Subsets

All the decision trees presented above used all the left-sided and midline variables collected in this study. However, time constraints usually prohibit such an exhaustive data collection process during the analysis of skeletal remains. In fact, most anthropologists would likely stick to those methods and definitions with which they feel comfortable. This author is no exception. In order to investigate the potential loss of information anthropologists could face by only choosing certain methods, several decision trees were constructed using certain subsets of the variables available. Three groups were used. First, decision trees were constructed using only variables from phase methods. These variables included phases of the pubic symphysis (Brooks and Suchey 1990; Todd 1920; Todd 1921), the auricular surface (Lovejoy et al. 1985b), and the sternal rib ends (İşcan et al. 1984a; İşcan et al. 1985). Second, decision trees were grown using only the component variables focusing on a single aspect of age indicators. These include the remainder of the other variables, amounting to 86 different variables spread over the four age indicators considered in this study: the pubic symphysis, the auricular surface, cranial sutures, and sternal rib ends. The last set of variables are those to which the author was exposed and taught to use during his collegiate career. These variables could be referred to as "standard" or "traditional" methods. In order to avoid ruffling too many feathers, hereafter this set of variables shall be referred to as "author's standards." These methods included all of the phase methods mentioned above as well as the component methods of McKern and Stewart (1957) and Gilbert and McKern (1973). For all three variable subsets, a CRT 30/10 and a CHAID 30/10 decision tree was constructed. For the phase method-only variables, both the decision trees were very simple using only three variables each. The CHAID 30/10 and CRT 30/10 decision trees using phase method-only variables are given in Figures 5.19 and 5.20 respectively.

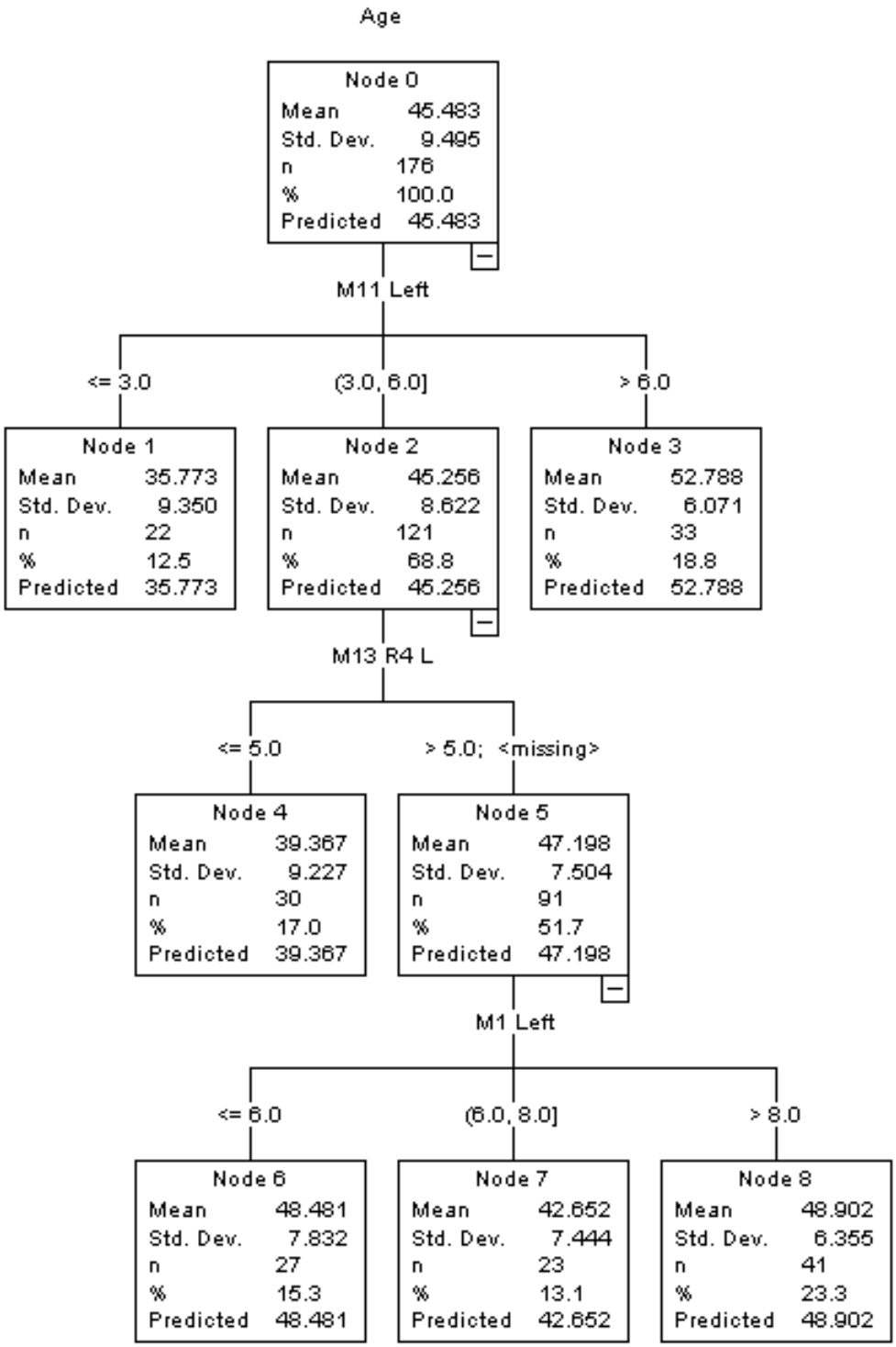


Figure 5.19: CHAID 30/10 with all observations and phase-only variables

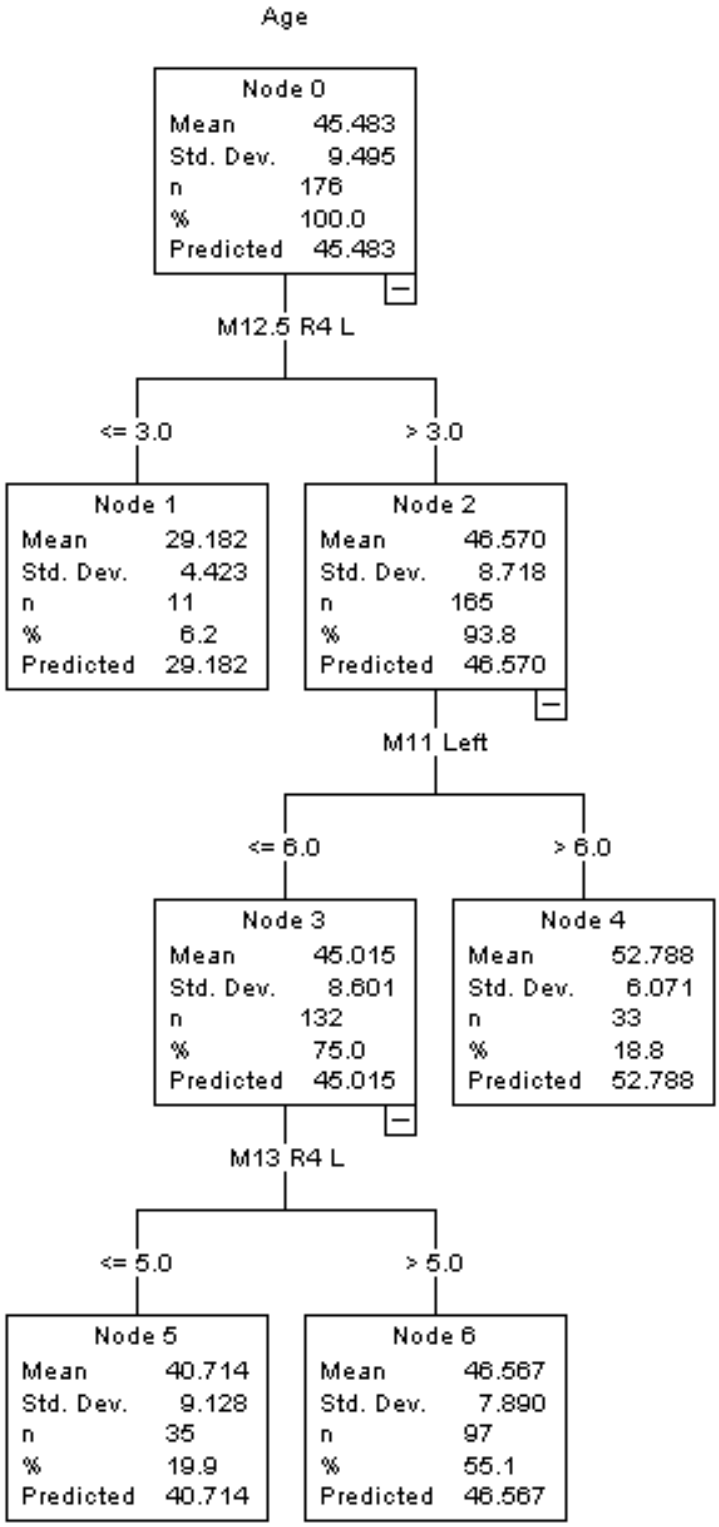


Figure 5.20: CRT 30/10 with all observations and phase-only variables

The variables selected for the CHAID 30/10 phase method-only tree (in order of importance) were the auricular surface (Lovejoy et al. 1985b), the female standard to the sternal rib end (İşcan et al. 1985), and the male standard of the Todd phase method for the pubic symphysis (Todd 1920). For the CRT 30/10 phase method-only decision tree the variables selected (in order of importance) were the sex-specific standards for rib end phases (İşcan et al. 1984; İşcan et al. 1985), the auricular surface (Lovejoy et al. 1985), and the female standard for sternal rib ends (İşcan et al. 1985).

These results bring up a couple of interesting points. First, the female standards for rib end phases and the auricular phase method show up in both trees and the pubic symphysis only shows up as the least influential variable in the CHAID decision tree. For these two trees, the growing algorithm had eight variables to choose from, four of which were related to the pubic symphysis. Of those four, three were closely related to each other, as they pertained to the male, female, and sex-specific standards for the phase methods put forth by Todd (1920; 1921). Similarly, three of the other variables pertain to the same set of standards related to the age changes in the sternal rib ends: male, female, and sex-specific (İşcan et al. 1984a; İşcan et al. 1985). Overall then, it is possible to argue that in reality only four variables were used: auricular phases (Lovejoy et al. 1985), rib phases (İşcan et al. 1984a; İşcan et al. 1985), and the pubic symphysis phase methods of Todd (1920; 1921) and Brooks and Suchey (1990). Either way one looks at the variables, half of them pertain to the pubic symphysis. If all of the variables in this distribution are equally valid, then we would expect about half of the variables in the decision tree to be related to the pubic symphysis. The two trees chose six variables, with only one being related to the pubic symphysis. This seems a little low and definitely suggests that the other phase methods are more applicable to estimating age than the pubic symphysis.

The second point of interest in these decision trees is the double occurrence of a rib method in the CRT tree (Figure 5.20). On first glance, this certainly appears strange, however, upon closer examination we realize that the variables are not quite the same and the splits with which they are associated are altogether different. One of the variables is the female standard and the other is the sex-specific standard. Considering that the data set is predominantly male, the sex-specific variable shares most of its values with the male variable. Therefore, the sex-specific variable and the female variable do not contribute exactly the same information. More importantly, in terms of the decision tree, even if the two variables were similar they are not separating the data set along the same split. The first split based on the sex-specific standard separates the data set into groups with phase expression of less than or equal to 3 and those with phases of greater than 3. On the other hand, the second split using the female definitions separates the observations in the parent node along phase 5. Taken together, these two variables could be seen as splitting the data into three subgroups with rib phases of three or less, phases 4 and 5, and phases greater than 5.

As these two trees are small compared to all of the other trees mentioned so far, it is easier to discuss specifically how these trees separate the data into age groups. In the CHAID 30/10 phase-only tree, the data is initially split into three groups: a younger, and older, and a middle group. The younger and older groups are defined solely by their expression of the auricular surface. The middle group, composed of individuals with auricular phases of 4, 5, and 6, are split then into two groups with a younger group being those of rib phases smaller than 6. Those individuals with auricular phases of 4, 5, or 6 and a rib phases greater than 5 are finally separated into three groups based on their pubic symphysis phases.

Since there are only binary splits in the CRT tree, it is slightly easier to look at. This tree has four age cohorts with the youngest being characterized by rib phases of 3 or less. The second youngest group has rib phases of 4 or 5 and an auricular phase of 6 or less. The third group also has an auricular phase of 6 or less but a rib phase of more than 5. The oldest group in this decision tree is characterized by rib phases greater than 3 and auricular surface phase greater than 6. Notably, in both the CRT and the CHAID tree, the split of the auricular surface between phases 5 and 6 and the split between phases 5 and 6 of the ribs were selected as the best splitting values.

The decision trees constructed using only component variables were larger and more complex than the phase only trees. This is to be expected if one considers these trees had 86 variables to choose from. The CHAID 30/10 diagram is presented in Figure 5.21 followed by the CRT 30/10 tree diagram in Figure 5.22. The CHAID tree uses seven variables to define nine distinct age groups. The CRT tree did not utilize pruning or surrogates and uses five variables to create six age groups. The variables selected by the CHAID decision tree growing algorithm were the pit shape of the sternal rib end (İşcan et al. 1984b), the degree of macroporosity on the auricular surface (Buckberry and Chamberlain 2002), the development of the ventral margin on the pubic symphysis (Boldsen et al. 2002), the development of the dorsal plateau of the pubic symphysis (Gilbert and McKern 1973), the formation of the dorsal edge of the pubic symphysis (Hanihara and Suzuki 1978), and the degree of suture closure at midlambdoid and sphenofrontal (Meindl and Lovejoy 1985). This CHAID decision tree is the only one of the simpler trees that uses a high number of variables related to the pubic symphysis.

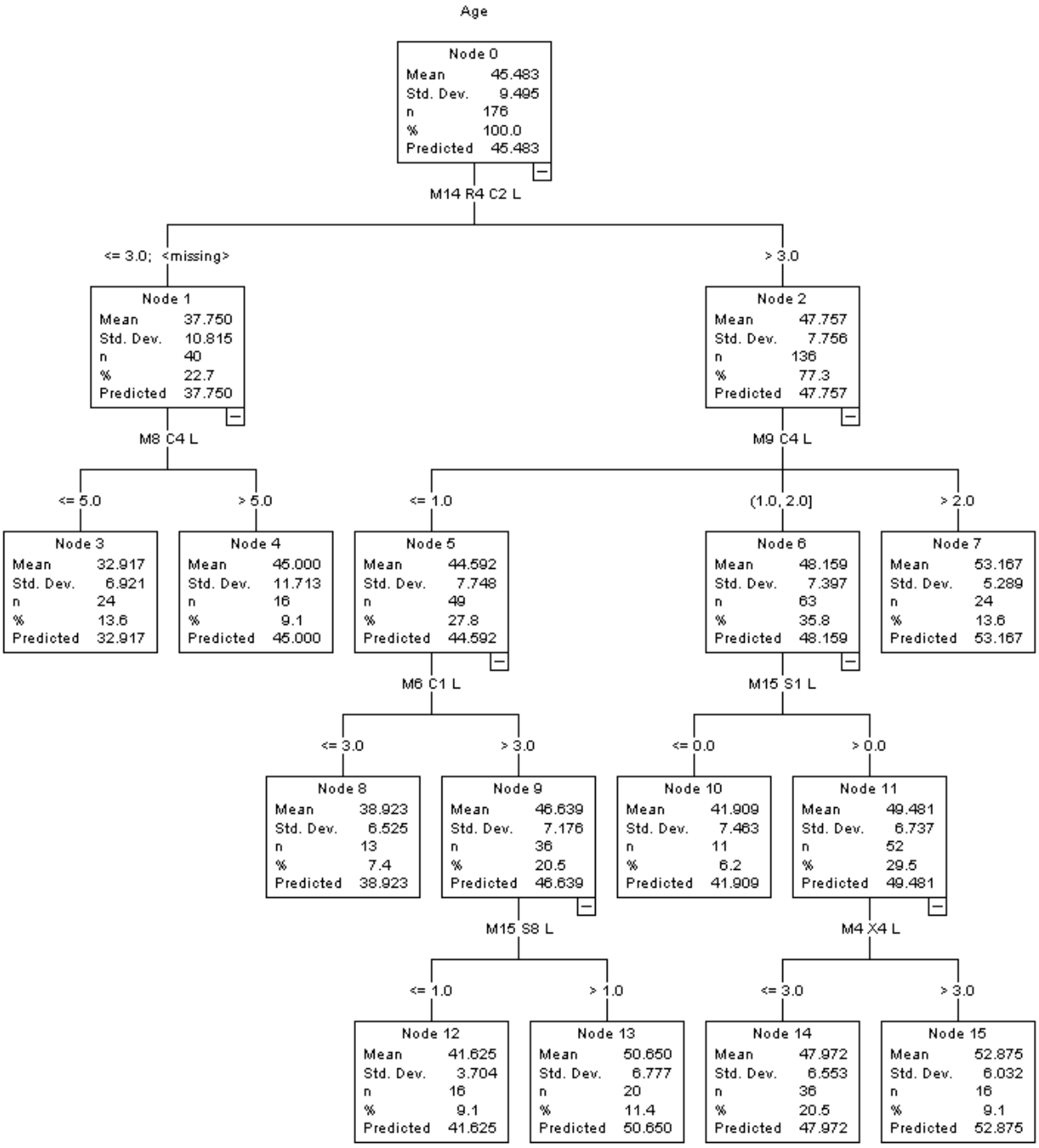


Figure 5.21: CHAID 30/10 with all observations and component-only variables

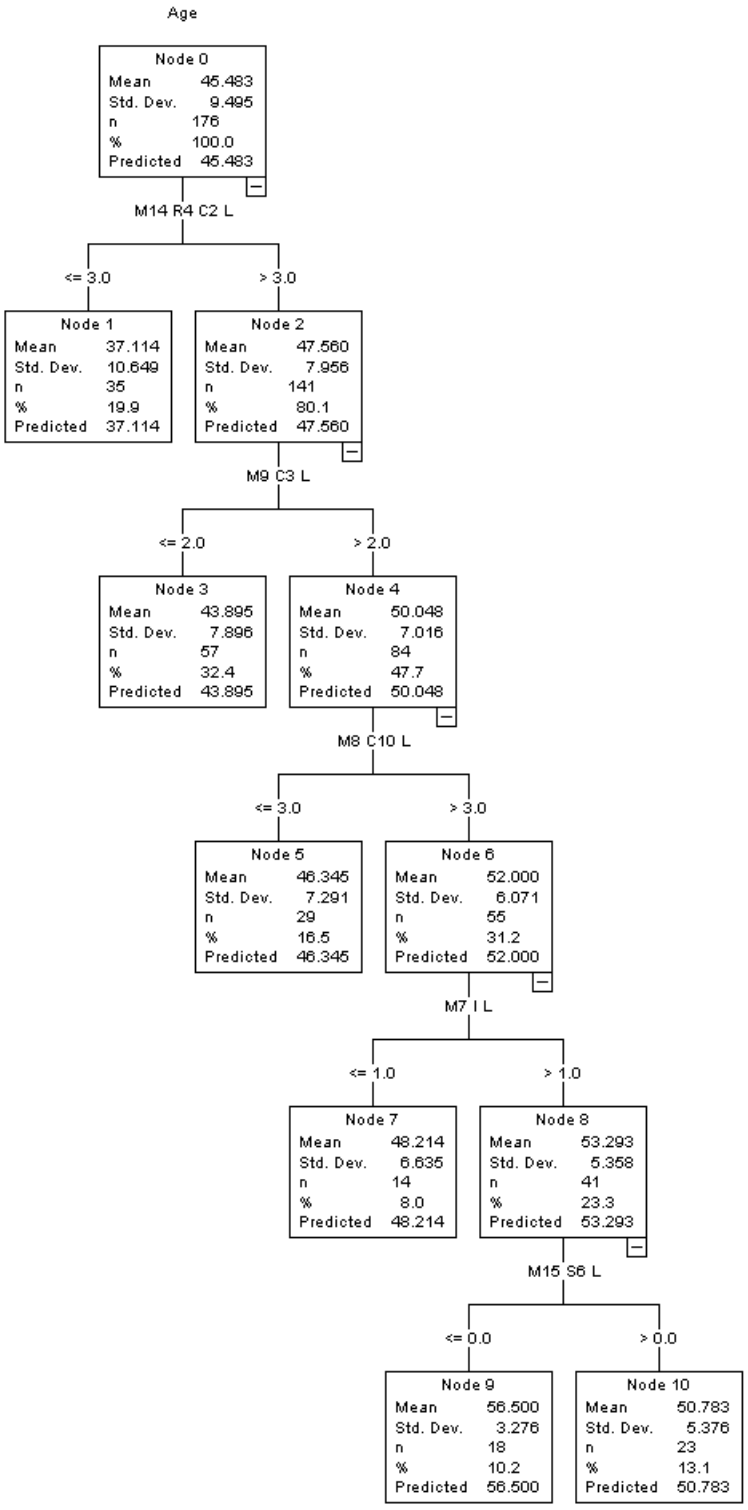


Figure 5.22: CRT 30/10 with all observations and component-only variables

The variables selected by the CRT tree growing algorithm were different. In particular these variables were the sternal rib end pit shape (İşcan et al. 1984b), the degree of microporosity on the auricular surface (Buckberry and Chamberlain 2002), the morphology of the inferior portion of the auricular surface (Boldsen et al. 2002), the surface texture of the pubic symphysis (Chen et al. 2008), and the degree of suture closure at the midcoronal point of observation (Meindl and Lovejoy 1985). Like almost all the other trees constructed in this study, this CRT tree does not consider the pubic symphysis as important an age indicator as the auricular surface and the rib morphology, which comprise the top three variables in this model.

The trees constructed using the author's standard variables are presented in the appendix in Figures C.7 and C.8. The CHAID decision tree is identical to the phase-only variable CHAID tree. It selects the same three variables which were (in order of importance) the auricular surface phase (Lovejoy et al. 1985b), the female phase standard to the sternal rib end (İşcan et al. 1985), and the male standard to the Todd phase method for the pubic symphysis (Todd 1920). The only difference between the author's standard variables and the phase-only variables is that the author's standards includes the McKern and Stewart (1953) and the Gilbert and McKern (1973) component variables for the pubic symphysis. The fact that the CHAID tree with phase variables is identical to the CHAID tree using the author's standards implies that these two component methods do not add any additional significant information to the model that the three variables in the model do not already provide.

The CRT decision tree given in appendix Figure C.8 shows a simple tree. It is similar to the phase-only variables CRT tree, but it adds a fourth variable and split. Using four variables this tree splits the data into five age cohort groups. The variables selected were the sex-specific standards for rib end phases (İşcan et al.1984a; İşcan et al. 1985), the auricular surface (Lovejoy

et al. 1985b), the female standard for sternal rib ends (İşcan et al. 1985) and the development of the pubic ventral rampart based on the female methodology (Gilbert and McKern 1973). The additional variable of the pubic symphysis is used to divide the oldest group of individuals depending on whether the stage of the ventral rampart is greater than 4 or less than 5, with the larger values determining the older of the two groups. The younger age groups are still determined by their rib stages and auricular surface phases.

Table 5.3 gives the goodness-of-fit measurements for the six decision trees that were constructed with only a subset of the variables available as well as the comparable values of the decision trees using all variables. For the CHAID decision trees, the component-only tree not only greatly outperforms the phase-only and author's standards trees but also slightly outperforms the CHAID decision tree using all variables. This indicates that the CHAID algorithm gets the most information from component variables and that the phase methods do not contribute a lot of additional information.

Table 5.3: Performance Summary Values for variable subset decision trees.

Model	Mean Absolute Error	% Variance
CHAID 30/10 all variables, all observations	5.5672	47.05%
CHAID 30/10, phase variables only	6.0960	37.52%
CHAID 30/10, component variables only	5.3763	49.59%
CHAID 30/10, author's standard variables only	6.0960	37.52%
CRT 30/10, all variables, all observations	5.1850	55.59%
CRT 30/10, phase variables only	6.2384	35.82%
CRT 30/10, component variables only	6.1128	35.55%
CRT 30/10, author's standard variables only	6.1489	36.78%

The main CHAID 30/10 decision tree (Figure 5.4) uses the phase of the auricular surface as the main variable in the initial split. The fact that the component-only model performs as well indicates that variables can be eliminated from a tree and a new tree can be as good, if new variables are selected to pick up the slack. Besides not having the auricular phase variable in it, the CHAID 30/10 component-only model shares several other variables with the base CHAID 30/10 tree. Among the variables that show up in both trees are the pit shape of the ribs (İşcan et al. 1984b), the degree of macroporosity on the auricular surface (Buckberry and Chamberlain 2002), and suture closure at midlambdoid (Meindl and Lovejoy 1985). The first two of these play major roles in both trees, while the suture closure variable only plays a mediocre part.

For the CRT trees, all three trees using variable subsets are very similar to each other. They also perform significantly worse than the CRT tree utilizing all variables, with age estimates that are on average a whole year worse than those of the base tree. This pattern suggests that the CRT growing algorithm finds useful information for age estimation in both phase and component variables. The CRT trees built on variable subsets are also much simpler than CRT 30/10, utilizing fewer variables. As such it is not surprising that the models are not as good as CRT 30/10 that uses all variables.

Comparative Regression Analyses

In order to compare the results of the decision trees constructed in this study, a regression analysis was performed for each of the decision trees using the variables selected by them. For each regression analysis, the r-square value was obtained and the mean absolute error (MAE) was calculated. Table 5.4 summarizes these results. It provides the mean absolute error for all the

Table 5.4: Performance Summary Values for decision tree and regression models.

	Decision Trees		Regression	
Original Model	MAE (yrs)	% Variance	MAE (yrs)	R^2
CHAID 100/50 all vars, all obs	7.00	19.31%	6.86	0.238
CHAID 10/5 all vars, all obs	3.81	70.25%	5.56	0.443
CHAID 10/5 all vars, w/o obs 81 and 116	3.60	74.15%	5.56	0.443
CHAID 20/10 all vars, all obs	5.33	50.98%	5.99	0.363
CHAID 30/5 all vars, all obs	5.21	52.28%	5.83	0.410
CHAID 40/10 all vars, all obs	5.68	45.85%	6.14	0.336
CHAID 30/10 all vars, all obs	5.57	47.05%	6.06	0.347
CHAID 30/10 all vars, w/o obs 146	5.46	49.84%	6.06	0.347
CHAID 30/10 all vars, w/o obs 10, and 146	5.14	53.68%	6.05	0.347
CHAID 30/10 all vars, w/o obs 10, 139, 146	5.55	46.78%	6.17	0.364
CRT 10/5 all vars, all obs	4.45	67.87%	5.28	0.505
CRT 10/5 all vars, w/o obs 15	5.36	50.01%	5.88	0.413
CRT 50/10, all vars, all obs	5.34	54.18%	5.65	0.441
CRT 30/10, all vars, all obs	5.19	55.59%	5.65	0.441
CRT 30/10, all vars, w/o obs 10	5.67	40.43%	6.03	0.365
CRT 30/10, all vars, w/o obs 10, 25	5.65	40.26%	6.03	0.365
CHAID 30/10, phase variables only	6.10	37.52%	6.28	0.338
CHAID 30/10, component variables only	5.38	49.59%	6.16	0.359
CHAID 30/10, author's standard vars only	6.10	37.52%	6.28	0.338
CRT 30/10, phase variables only	6.24	35.82%	6.23	0.354
CRT 30/10, component variables only	6.11	35.55%	6.40	0.301
CRT 30/10, author's standard variables only	6.15	36.78%	6.23	0.354

decision trees, the percentage of the variance in age that was accounted for by the tree, and the comparable values obtained from the regression analyses.

The performance of the multiple regression is worse for all but two of the models. While the percentage of the variance in age accounted for by the decision trees is a measure of the strength of the model, it is not directly comparable to the R^2 values of the linear regressions. These values are better used to compare the strength of different decision trees to each other. However, the mean absolute error values are based on the same information and can be used to

compare the two model types. This pattern suggests that while linear regression can detect the general pattern of age estimation with the age indicators defined by anthropologists, the actual mechanism of aging, or the way the age indicator stages are defined, is not something that is grasped by regression analysis. Furthermore, using the same variables, decision trees are able to come up with more accurate models than regression analysis can. This can likely be expressed by how decision trees function. They only focus on a single split in one variable at a time and also apply the splits they find as significant only to the subsets of data in the parent node they are attempting to split. Regression, on the other hand, has to use the whole variable and apply it to the whole data set. Hence, if a variable only has a strong correlation with part of the data but does not apply equally well to the other parts, those parts of the data to which the variable does not relate drag down the overall performance of the variable in the regression analysis. In addition, if an ordinal variable (as most age estimation variables are) is well defined for only some of its ordinal values, the same drag on the overall performance of the variable would be felt in the regression analysis. The fact that the decision trees slightly outperform regression suggests that the way age indicator stages relate to age are not as straightforward as regression assumes them to be. It further suggests that specific age indicator stages are more important than others.

The two instances when the regression analysis exceeded or came close to the decision tree results were for CHAID 100/50 and for CRT 30/10 phase variables only. CHAID 100/50 was the default tree and only used two variables. It was not really a great decision tree due to its limited size. Nevertheless, using the same two variables the regression analysis did better, although not by an overwhelming amount. The closeness in performance between CRT 30/10 phase-only and the corresponding regression analysis is likely due to the small number of variables. Fewer variables are always worse at predicting age than more, since more variables

obviously contain more data. The three variables selected for the CRT 30/10 phase-only decision tree contain little discriminating information and therefore regression using the same variables performs just as well. This suggests that decision trees get better with size, at least with respect to comparative regression analyses. In terms of age estimation, it suggests that the basic patterns can be defined by a regression equation, but that there are more complex patterns among the age indicators that the decision trees can ferret out once they are allowed to grow beyond a few splits.

Another interesting point from Table 5.4 is that the large decision trees had much better performance than regression equations that used the same variables as those trees. The CHAID 10/5 decision trees did especially well compared to the regressions. The question becomes, with both models using the same information, how can one fare so much better than the other? The answer is that decision trees are more adept at finding the valuable information pertaining to age that is contained in a very large number of aging methods and age indicators.

The final analysis was a stepwise variable selection regression analysis. With a 0.05 level of significance required to be entered into the equation, the regression analysis picked 13 variables out of the 94 from which it had to choose. The variables selected are in order:

1.) the age-specific phases of the rib (İşcan et al. 1984a; İşcan et al. 1985)

2.) the phases of the auricular surface (Lovejoy et al. 1985)

3.) suture closure at the incisive suture (Meindl and Lovejoy 1985)

4.) suture closure at the sphenofrontal (Meindl and Lovejoy 1985)

5.) the amount of superior posterior iliac exostoses (Boldsen et al. 2002)

6.) the development of the dorsal margin of the pubic symphysis (Boldsen et al. 2002)

7.) the sex-specific ventral bevel development of the pubic symphysis (McKern and Stewart 1953; Gilbert and McKern 1973)

8.) the dorsal aspect of the pubic symphysis (Hanihara and Suzuki 1978)

9.) the degree of billowing at the apex of the auricular surface (Boldsen et al. 2002)

10.) the development of the ventral bevel of the pubic symphysis (Boldsen et al. 2002)

11.) the presence of coarse granularity on the auricular surface (Igarashi et al. 2005)

12.) suture closure of superior sphenotemporal (Meindl and Lovejoy 1985)

13.) the amount of posterior iliac exostoses as defined by this author derived from Boldsen and colleagues (2002)

This regression equation accounts for about 60.8% of the variance in the data and has a mean absolute error of estimation of 4.69 years. These numbers are quite good when compared to those of the decision trees. They are similar to the numbers of CHAID 10/5 and CRT 10/5, which were the maximum decision trees grown and considered to be overfitting the model to the data. Likewise, the 13 variable regression equation could be considered too cumbersome to be of practical use.

Figure 5.23 shows a graph of the predicted ages against the actual ages for this regression equation. We see a nice even scatter around the true age line. We further see what was mentioned as a problem with regression in the literature review, that the young individuals are generally overages, while the old individuals are underaged.

When we reduce the number of variables used in the regression equation and only use the top eight variables (i.e., all variables up to and including the dorsal development of the pubic symphysis as defined by Hanihara and Suzuki in 1978), the explained variance in age proportion drops to 53.8% and the mean absolute error of the predictions is 5.21 years. These numbers are

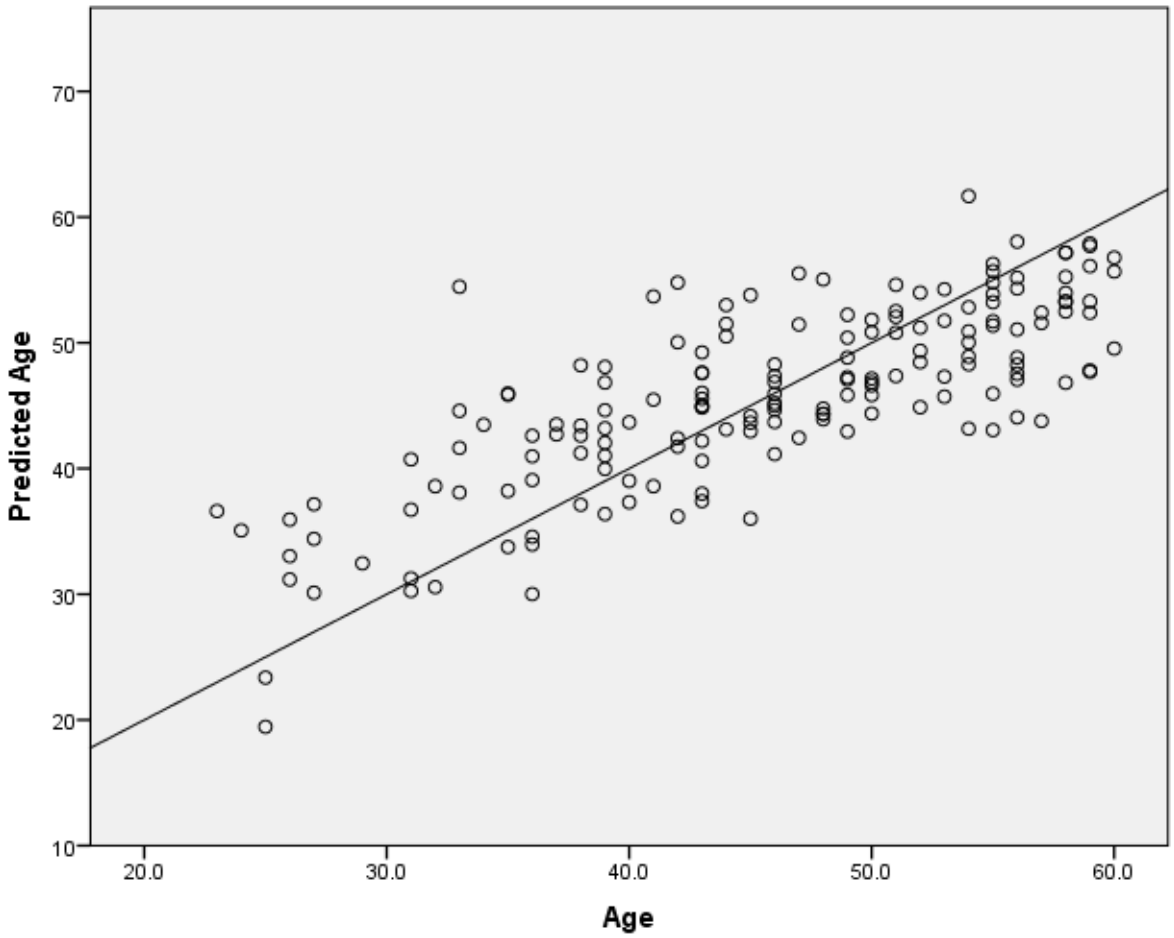


Figure 5.23: Predicted ages for the 13 variable stepwise regression plotted against actual ages

much more comparable to those of the more appropriate decision trees, such as CRT 30/10 and CHAID 30/10 with observations 5, 10, and 146 removed. Further, the number of variables used in this regression equation is similar to the number of variables that were chosen in most of the medium sized decision trees. This reduced stepwise regression model performs very well compared to the decision trees. Nevertheless, this author stipulates that there is still information pertaining to age estimation in the age indicators that decision trees are better designed to detect.

Results Summary

All the decision trees mentioned above, like most other statistical methods as well, are in one way or another a measure of how well related the various age indicator states are to recorded actual age. One would expect that if data is collected on 94 different variables, a relationship would be found between morphology of an age indicator and actual age. However, no anthropologist wants (or has the time) to collect 94 points of data simply for a single age estimate. The primary purpose of this research was to find out which age indicator variables, of all those possible, are good variables for age estimation.

In order to determine which variables were the most valuable in this particular study, the number of times each variable was utilized in a decision tree was counted. As several of the decision trees are similar to each other (or at least derivatives of one another), not all decision trees grown in this study were utilized in this variable accumulation. In total nine decision trees were used. These were judged to represent different sets of variables that, when combined, would not weigh a single pattern multiple times. The decision trees that were used were CHAID 10/5, CHAID 30/10, CHAID 30/10 without observations 10, 139, and 146, CRT 10/5, CRT 10/5 without observation 15. Two versions of the CRT 10/5 tree were included because each was deemed different enough and presented a different subset of variables used in the analysis that they would not present double counting and cause overemphasis of its variables. The same justification applies to the selection of two CHAID 30/10 trees. Furthermore, the CHAID and CRT phase-only and component-only trees were also included as they represent complementary sets of variables. None of the CRT 30/10 decision trees were included because they all represented a smaller version of one of the two CRT 10/5 trees and as such contained the same set of variables. In addition to the CRT and CHAID trees mentioned above, the 13 variables from

the stepwise regression analysis were also included in the summation. Summarily, the variables of ten models were combined. However, because the phase-only and component-only trees represent complimentary sets of variables, a single variable could only appear a maximum of two times in these four decision trees. Therefore, over all ten models the maximum number of appearances for any one variable is eight.

Table 5.5 shows the counts of the eleven most common variables encountered during the analyses. The table is abridged for brevity, as there are eleven more variables that occurred twice in the models selected and 15 variables that occurred just once in the different models. The most commonly selected variables were the auricular surface phase variable (Lovejoy et al. 1985b) and the pit shape of the sternal rib ends (İşcan et al. 1984b). Reviewing the various decision trees, these two variables are generally found toward the top of the tree as well, indicating their importance. Variables near the top of the tree are more influential as they apply their classification rules to a greater number of observations. In contrast, variables further down the decision trees exhibit less influence over the data, as they apply their rules to a smaller number of

Table 5.5: Most common variables used in distinct decision trees and regression models.

Code	Variable	Bone	Type	Counts
M11	Auricular surface	Auricular	Phase	6
M14 R4 C2	Pit shape	Ribs	Component	6
M9 C4	Degree of Macroporosity	Auricular	Component	5
M15 S1	Midlambdoid closure	Sutures	Component	5
M4 X4	Dorsal Margin	Pubis	Component	4
M12.5 R4	Sex-specific rib morphology	Ribs	Phase	4
M15 S6	Midcoronal closure	Sutures	Component	4
M15 S8	Sphenofrontal closure	Sutures	Component	4
M8 C10	Inferior surface morphology	Auricular	Component	3
M9 C2	Surface texture	Auricular	Component	3
M8 C4	Ventral margin	Pubis	Component	3

observations. The two most common variables are also the first two variables that were selected by the stepwise regression analysis, further highlighting their strong correlation to age in the data set. The amount of macroporosity on the auricular surface (Buckberry and Chamberlain 2002) and the degree of suture closure at midlambdoid (Meindl and Lovejoy 1985) occur second most frequently among the ten models aggregated. These two variables are found further down the decision trees but are generally still part of the core. Neither of these two variables was chosen in the stepwise regression equation, however, a disparity suggesting that the decision tree algorithm saw something in these variables that the regression did not pick up on. It is possible that these variables were only good at distinguishing based on a couple of their defined age stages, rather than all of them; therefore being overlooked by the regression analysis and picked by the decision tree precisely for their specific distinguishing capabilities.

Table 5.6 lists the number of variables in each site category that were in the data set, combining the various variables into categories depending on which age determination site the variable pertains to. In this research only the four most common age estimation sites were utilized: the auricular surface, the sternal rib ends, the pubic symphysis, and cranial sutures. The third column of the table provides the number (and proportion) of variables that occurred at least once in the ten models selected above for the variable summation for that aging site. The fourth column lists the number of times any variable of a particular age estimation site occurred in the set of models. The final column shows the average number of times a variable selected to be in a model showed up in the eight models that were selected for variable summation.

Table 5.6: Breakdown of variable occurrence by aging site.

Age Indicator Sites	Number of Variables	Variables in Models	Number of Occurrences	Average Number of Times used
Auricular Variables	30	14 (47%)	32	2.3
Rib Variables	6	4 (67%)	14	3.5
Pubic Symphysis Variables	39	11 (28%)	19	1.7
Suture Variables	19	8 (42%)	19	2.4
Total	94	37 (39%)	84	

The results show that the highest proportion of variables selected pertained to age-related changes of the rib. Furthermore, about half of all the auricular and suture variables were utilized at some point during the analysis. By comparison, less than one third of all the pubic symphysis variables were chosen during the analyses. These percentages could certainly be an artifact of the quality of the age indicator variables used. It is possible that some of the 39 pubic symphysis variables are just bad and show no relationship to age at all. However, we would expect all age indicator sites to have about the same amount of bad variables, so the percentages are not without meaning. The table also shows that the auricular variables, pubic symphysis variables, and suture variables occurred on average around two times each in the set of models selected. However, the pubic symphysis was chosen less often than both the auricular surface and suture variables, and less than half as often as the rib variables which averaged more than three occurrences. Age-related variables of the ribs were very common in the analysis and as such would suggest that they are worthwhile to consider for age estimation methodologies. Overall, the table indicates that pubic symphysis variables underperformed in these analysis and the rib variables performed exceptionally well.

Jackknifed Decision Trees: CHAID

The basic CHAID 30/10 decision tree with all observations is given in Figure 5.4. It has an MAE of 5.57 years and the model is focused on three variables of the auricular surface and a single variable of the ribs, with a couple of suture closure variables at the bottom of the tree. The top four variables in this tree are the phase of the auricular surface (Lovejoy et al. 1985b), the surface texture and the amount of macroporosity on the auricular surface (Buckberry and Chamberlain 2002), and the pit shape of the sternal rib ends (İşcan et al. 1984b).

The six jackknifed CHAID 30/10 decision tree diagrams are given in the appendix in Figures C.9 through C.14. The CHAID decision tree without age group 1 shown in Figure C.9 illustrates what happens when a narrower age range is investigated. Having eliminated ages 23 to 35 from this analysis, this tree pertains to the older group of individuals in the data. The tree selects the amount of macroporosity on the auricular surface (Buckberry and Chamberlain 2002), the surface morphology of the inferior portion of the auricular surface (Boldsen et al. 2002), and the suture closure at pterion, sphenofrontal, and of the incisive suture (Meindl and Lovejoy 1985). The auricular surface variables separate the data into the oldest and the youngest groups while the suture closures are used to distinguish between smaller subgroups of the data. In terms of performance, the MAE for the training sample (i.e., ages 36 to 60) is better with only an error of 4.6 years. This is due to the fact that the data used for this decision tree is narrower and thus contains less diversity. By reducing the distribution of the age, the models are naturally going to look better. Additionally, leaving out observations at one end of the age spectrum requires the model to extrapolate ages for those individuals in the jackknifed age group. Therefore, age estimates for the jackknifed individuals are expected to be very high since decision trees and many other statistical models have problems extrapolating estimates outside the ranges used to

construct the models. As expected, the MAE for the jackknife age group of individuals ages 23 to 35 is very high. With a jackknife MAE of 15.2 years, this decision tree definitely does not work well for individuals between the ages of 23 and 35.

The CHAID tree without age group 2 presented in Figure C.10 shows an interesting pattern. With individuals used in the age ranges 23 to 35 and 41 to 60, the decision tree shows a pattern that separates these different age ranges within the tree. The decision tree uses the inferior edge of the pubic symphysis as defined by this author (based on Hanihara and Suzuki 1978) to separate 15 young observations out of the data. While that only represents about half of the observations in the 23 to 35 age range, it nevertheless shows the importance of the pubic symphysis for age estimation in younger individuals. After the split of the data into a younger and an older subset, the decision tree uses the phase of the auricular surface (Lovejoy et al. 1985b), the roughness (Igarashi et al. 2005), and the surface texture (Buckberry and Chamberlain 2002) of the auricular surface to break up the older portion of the data set. Since this jackknifed data sample takes a chunk of data out of the middle of the age range it does not narrow the age range. Therefore, the mean absolute error values are more telling about the strength of this decision tree than the one without age group 1. With an MAE on the training sample of about 6 years, the tree performs about as well as the tree using all observations. The MAE of the held out observations is about 8 years, which is better than that of the age group 1 jackknifed tree, but a little higher than we would want.

The CHAID tree without age group 3 is shown in Figure C.11. This tree excludes observations between the ages of 41 and 45 and performs better than either of the previous two jackknifed decision tree. It also shows some important similarities to the overall CHAID 30/10 tree. It uses the phase of the auricular surface (Lovejoy et al. 1985b) and the pit shape of the

sternal rib ends (İşcan et al. 1984b) as the most influential variables. Both of these variables are also the primary variables used in the overall CHAID tree, and both trees also use the same cutting points for these variables. The auricular phase is separated at stages 3 and 6 and the pit shape is separated at category 3 in both trees. Less influential variables in this decision tree are the dorsal edge of the pubic symphysis (Hanihara and Suzuki 1978), closure of the superior sphenotemporal suture (Meindl and Lovejoy 1985), and the presence of dense porosity on the auricular surface (Igarashi et al. 2005). With an MAE of 5.6 years on the training sample and a comparable MAE of 6.6 years on the jackknifed age group, this decision tree performs the best out of all the jackknifed trees seen so far.

The CHAID decision tree without age group 4, as seen in Figure C.12, is one of the simpler decision trees we have seen. It only uses the phase of the auricular surface (Lovejoy et al. 1985) and the degree of suture closure at pterion (Meindl and Lovejoy 1985). The tree uses the same splitting criteria (3 and 6) of the auricular phase as we have seen in several other decision trees and splits the data into a younger, an older, and a middle age category. The suture closure then splits the middle age group into three more specifically defined age groups. Even though this decision tree is simpler than the other CHAID trees that have been grown it actually performs on par with the other trees with a training sample MAE of 6.7 years. What is of particular interest is that the jackknife MAE is only 4.27 years, as this is by far the smallest jackknife MAE that any of the decision trees show. This small value indicates that this simple decision tree is very good at approximating the age of individuals that fall in the middle of its age range.

Exclusion of age group 5 yields a CHAID decision tree that shares very few traits with the other trees. Seen in Figure C.13, this tree selects the pit shape of the sternal rib end (İşcan et

al. 1984b) with category 3 as the splitting category of the primary variable to split the data set into a smaller younger age group and the larger group of older individuals. Rib pit depth (İşcan et al. 1984b) and surface appearance of the pubic symphysis (Chen et al. 2008) are used to more closely define younger age groups. The degree of microporosity on the auricular surface (Buckberry and Chamberlain 2002), the surface morphology of the inferior portion of the auricular surface (Boldsen et al. 2002), and suture closure at midcoronal (Meindl and Lovejoy 1985) are used to further divide the older age group into smaller portions. The training MAE for this tree of 5.7 years is comparable to the other trees, while the jackknife MAE of 8.9 years is a lot higher than the decision trees that exclude age groups 3 and 4. Since this tree is not as accurate at estimating age of individuals for the group of individuals that were held out, this tree may be more training sample specific than we would like for age prediction of other unknown individuals.

The final CHAID jackknife tree excludes age group 6 and therefore is built based on individuals ages 23 to 55. Seen in Figure C.14, this tree shows the same characteristics as the first decision tree that excluded age group 1. This is again due to the reduced age range used in the construction of this decision tree and the need for the tree to extrapolate age estimates for the jackknifed individuals, as they fall outside the age range from which the tree was designed. The tree shows a very high jackknife MAE of 12.2 years due to the extrapolation required to estimate the ages of the jackknife subsample. The training sample MAE is very small at 4.9 years and is again likely due to the smaller age range from which this tree is constructed. The variables selected for this tree are interesting as each subsequent variable factors a younger portion of individuals out of the sample. The youngest subgroup is defined by the inferior edge of the pubic symphysis as defined by this author (based on Hanihara and Suzuki 1978). The second variable

is the phase of the rib morphology (İşcan et al. 1984a) and factors out a small and slightly older age group from the sample. The transverse organization and the amount of macroporosity on the auricular surface (Buckberry and Chamberlain 2002) and the degree of suture closure at midlambdoid (Meindl and Lovejoy 1985) define further subgroups for the older individuals in the sample.

Table 5.7 summarizes the various MAE values for the CHAID decision trees with the different age groups left out of the analysis. In terms of performance, the best decision tree out of this set is the one with age group 4 left out, followed by the one without age group 3. Both of these decision trees show that they are able to predict ages for individuals that are not included in the age range used to construct the decision tree. The tree without age group 4 indicates that a simpler tree may be more useful in age estimation. Both of these decision trees highlight that the phase of the auricular surface (M11 Left) is very useful in the age estimation process.

Table 5.7: Age ranges and mean absolute error values for jackknife CHAID decision trees

Decision Tree	Age Ranges	MAE_{jack}	MAE_{train}	MAE_{total}
CHAID 30/10	23-60			5.57
Without age group 1	36-60	15.20	4.62	6.30
Without age group 2	23-35 & 41-60	8.00	6.00	6.28
Without age group 3	23-40 & 46-60	6.59	5.60	5.77
Without age group 4	23-45 & 51-60	4.27	6.74	6.29
Without age group 5	23-50 & 56-60	8.92	5.69	6.24
Without age group 6	23-55	12.20	4.95	6.22

Jackknifed Decision Trees: CRT

The basic CRT 30/10 decision tree with all observations is given in Figure 5.16. It has a MAE of 5.19 years and since CRT decision trees only use binary splits they are naturally narrower and deeper than CHAID decision trees, which can have multi-way splits. The CRT tree using all observations uses sex-specific rib phases (İşcan et al. 1984a; İşcan et al. 1985), the phase of the auricular surface (Lovejoy et al. 1985b), the surface texture and the amount of macroporosity on the auricular surface (Buckberry and Chamberlain 2002), and suture closures at midcoronal, anterior sagittal, pterion, and superior sphenotemporal (Meindl and Lovejoy 1985). The rib morphology and the auricular surface texture are used to separate out the younger group of individuals. The overall mean absolute error of the complete CRT 30/10 decision tree is 5.18 years.

The jackknife CRT 30/10 decision trees are given in the appendix in Figures C.15 through C.20. The decision tree excluding age group 1, shown in Figure C.15, is very different from most of the other trees that we have looked at. It uses the development of the ventral pubic margin (Boldsen et al. 2002) to define the youngest group of individuals in the data set. Two auricular surface variables are more important for older individuals. These are the surface morphology of the inferior surface (Boldsen et al. 2002) and the amount of observable microporosity (Buckberry and Chamberlain). Suture closure at pterion and sphenofrontal are used by the tree to define further subgroups farther down the tree. This tree shows the type of decision tree or statistical model that could be useful for only a portion of the age continuum. There is no auricular phase variable or any variable pertaining to the morphological changes observed in the rib, showing that if you are only interested in an older portion of the population the variables that are significant differ from those when you are looking at the whole age range.

The mean absolute error for this decision tree with the excluded age group 1 follows the pattern established by the CHAID decision trees. Due to the need to extrapolate age estimates for the jackknife observations, the jackknife MAE is very high at 15.7 years, and the narrower age range used to build the model by excluding age group 1 yields a lower training sample MAE value of 4.7 years.

The CRT decision tree with age group 2 excluded is given in Figure C.16. It shows a tree that is similar to the overall CRT tree that used all observations. It is actually a CRT 30/8 decision tree, because the young age groups were too small to be differentiated by a CRT 30/10 decision tree. Its first four variables are the same four variables in the same order and using the same splitting categories as the overall CRT tree. It uses the sex-specific rib phases (İşcan et al. 1984a; İşcan et al. 1985), and the surface texture of the auricular surface (Buckberry and Chamberlain 2002) to define the two youngest groups in the sample. It further uses the same split for the phase of the auricular surface (Lovejoy et al. 1985b) and suture closures at pterion (Meindl and Lovejoy 1985). Lastly, it used the development of the dorsal pubic margin (Hanihara and Suzuki 1978) to distinguish between individuals aged in the low 40s and high 40s. The training sample MAE of 5.4 is similar to that of the overall tree, which is slightly more complex and therefore is expected to perform better. The MAE for the jackknifed age group (ages 36 to 40) is slightly higher than others at 7.3 years indicating that this decision tree does not pick up on an aging pattern that would be useful for aging the excluded individuals.

The CRT decision tree excluding age group 3 is given in Figure C.17. It is a relatively simple tree that performs much like the other jackknife CRT trees. For this decision tree the ages used were 23 to 40 and 46 to 60. The decision tree used the pit shape of the rib end (İşcan et al. 1984b) to separate out the youngest 32 individuals in the data sample. The tree then uses the

degree of microporosity on the auricular surface (Buckberry and Chamberlain 2002) and the suture closure at sphenofrontal (Meindl and Lovejoy 1985) to divide the rest of the sample. In total this decision tree only has 4 terminal age groups. The training MAE is 6.2 years which is slightly higher than is generally seen and can be ascribed to the simplicity of the tree. The MAE of the jackknife sample is only slightly larger with about 6.6 years, suggesting that this model is good at estimating age for the whole age range, even though it did not use any individuals of ages 41 to 45 to come up with the age classification criteria.

The CRT decision tree without individuals in age group 4 is shown in Figure C.18. It shows the best performance of age estimation for the jackknife group with an MAE of 5.6 years. The training MAE of 6.5 years is nothing special. This decision tree uses the pit shape of the sternal rib ends (İşcan et al. 1984b) to define the youngest group of individuals. The splitting criteria for this variable are the same as in the jackknife CRT tree without age group 3. Individuals with a score of 3 or less are placed in the younger group and those with higher scores move into the older category where they are further differentiated by the degree of microporosity observable on the auricular surface (Buckberry and Chamberlain), the surface morphology of the inferior portion of the auricular surface (Boldsen et al. 2002), and the degree of suture closure at midcoronal (Meindl and Lovejoy 1985). The splitting criteria for the degree of microporosity are also the same as that in the tree without age group 3.

The CRT decision tree without age group 5 is given in Figure C.19. It may look familiar because it is the exact same tree as that when age group 4 was excluded from the analysis. It has the same variables, in the same order, with the same splitting criteria. The training MAE for this tree is slightly lower than the previous tree with a value of 5.9. Contrarily, the jackknife MAE is higher with a value of 8.8 years, indicating that this particular decision tree does not classify

individuals in age group 5 (age 51-55) as well as those in age group 4 (ages 46-50). Furthermore, three of the decision trees have two variables with the same splitting category in common and place those variables at the first and second branch of the decision tree. Pit shape of the rib end and the amount of microporosity show up in the CRT decision trees that exclude age groups 3, 4, and 5. Since all three of these trees were built with a data sample that had a portion taken out of the middle of the age range, we can infer that these two variables have significant separation power between older and younger individuals.

The final jackknife CRT decision tree excludes age group 6, individuals of ages 56 to 60. The tree, given in Figure C.20, represents the "younger" portion of the data set. The decision tree uses the male rib morphology phases (İşcan et al. 1984a), the development of the inferior edge and the development of the ventral bevel of the pubic symphysis as defined by this author (based on Hanihara and Suzuki 1978), the amount of transverse organization on the auricular surface (Buckberry and Chamberlain 2002), and the degree of suture closure of the incisive suture (Meindl and Lovejoy 1985). These variables represent features observed earlier in the aging process, and without the old individuals in the sample, these variables show more significance than they would if older individuals were included in the analysis. The mean absolute error for this decision tree is once again not helpful because it is influenced primarily by the data used in this decision tree. The jackknife age group is outside the age range of the decision tree and shows a corresponding high MAE value of 12.3 years, and due to the narrower age range of the data used to grow this tree the training MAE is expectedly lower with a value of 4.8 years.

Table 5.8 summarizes the various mean absolute error values for the jackknifed CRT decision trees. It shows that the CRT tree without age group 4 has the best performance on the

Table 5.8: Age ranges and mean absolute error values for jackknife CRT decision trees

Decision Tree	Age Ranges	MAE_{jack}	MAE_{train}	MAE_{total}
CRT 30/10	23-60			5.19
Without age group 1	36-60	15.68	4.68	6.43
Without age group 2	23-35 & 41-60	7.27	5.43	5.69
Without age group 3	23-40 & 46-60	6.55	6.19	6.25
Without age group 4	23-45 & 51-60	5.56	6.50	6.33
Without age group 5	23-50 & 56-60	8.79	5.87	6.37
Without age group 6	23-55	12.26	4.79	6.10

jackknife sample, followed by the tree excluding age group 3. Both of these trees are similar and also similar to the tree without age group 5 as mentioned earlier.

While the similar decision trees and variables in the middle jackknife samples show that there are general trees that perform well for the whole age range, the fact that the tree structures and variables change frequently between jackknife samples indicates that different variables work better for different sections of the age distribution. Therefore, it may be worthwhile to consider models that are able to apply different variables to different portions of the age continuum.

Logistic Regression

Several logistic regressions were run on the data sample as well. The whole data set was used as well as each of the six jackknifed data subsamples. The focus was primarily on the variables selected by the method and how those variables changed from jackknifed subsample to subsample. For the logistic regression, age was treated as a categorical variable and the same age groups from Table 4.1 were used. Further, a CRT 30/10 and CHAID 30/10 decision tree was grown using the same age groups as categories to provide a comparable measure of performance in the percentage of observations classified correctly. Table 5.9 summarizes these models.

Based on the numbers, the logistic regressions that excluded age groups 1 and 2 performed the best. These models also used more variables which may contribute to their superior performance. In terms of correct classification both of the decision trees perform worse than any of the logistic regressions. However, the classification of the logistic regression models is not great, with correct classification values around 65%. Since we already removed some of the age estimation precision of the model by converting age to categorical intervals, we would really want the classification of the logistic regressions to be better. One reason for the low classification percentages is that the correct variables were not used in the analysis. Looking at interaction variables between two or more age indicators would potentially increase the performance of the logistic regression. However, with 94 different variables there are a lot of potential interactions to consider. Interaction variables were not incorporated in these analyses, but should be considered for future analyses.

Table 5.9: Performance Summaries of Logistic Regression and Decision Trees using Age Categories

Decision Tree	Age Ranges	# of Variables	% Classified Correctly
Logistic Regression	23-60	6	50.9%
Without age group 1	36-60	9	69.5%
Without age group 2	23-35 & 41-60	8	72.7%
Without age group 3	23-40 & 46-60	4	55.5%
Without age group 4	23-45 & 51-60	7	66.4%
Without age group 5	23-50 & 56-60	6	63.0%
Without age group 6	23-55	4	58.5%
CHAID 30/10	23-60	8	46.6%
CRT 30/10	23-60	5	39.8%

Table 5.10 lists the variables that were entered into the different logistic regressions. Variables that appear three times or more are highlighted. The most common variable was the phase of the auricular surface (M11) (Lovejoy et al. 1985b). Other repeat variables were the pit shape of the rib end (M14R4C2) (İşcan et al. 1984b), the surface morphology in the apical region of the auricular surface (M8C9) (Boldsen et al. 2002), the amount of microporosity observable on the auricular surface (M9C3) (Buckberry and Chamberlain), and the degree of closure of the superior sphenotemporal suture (M15S10) (Meindl and Lovejoy 1985).

The table is very confusing, and while it is hard to see any patterns in the variables selected at all, a couple of patterns do seem to be present. First of all, most of the repeat variables that are highlighted show up in the upper half of the variables. This means that for the decision trees, these variables have more influence over the data set. For the logistic regression, the variables are listed in the order they were entered into the model, therefore the top variables represent variables with stronger links to age. Since the more important variables are the ones that appear more frequently across the models, these variables have strong ties to the age categories used in these analyses.

There are some patterns that can be observed in Table 5.10 as well with respect to the variables selected in individual models. The phase of the auricular surface (M11) does not show up in the logistic regressions without age groups 1 or 6. Since the variable only shows up in models that use the complete age range, regardless of which age groups are jackknifed, M11 looks like a variable that is useful to separate the young from the old. We have seen a similar use of M11 in various decision trees throughout these analyses. Furthermore, looking at the variables selected by the jackknife logistic regression without age group 1 (ages 23-35) shows that only two of the variables show up in at least two other models.

Table 5.10: List of Variables selected for Logistic Regressions and Categorical Decision Trees

Logistic Regression	Without Age Group 1	Without Age Group 2	Without Age Group 3	Without Age Group 4	Without Age Group 5	Without Age Group 6	CHAID 30/10	CRT 30/10
Ages 23-60	36-60	23-35/41-60	23-40/46-60	23-45/51-60	23-50/56-60	23-55	23-60	23-60
M11	M8C9	M14R4C2	M11	M11	M11	M14R4C2	M14R4C2	M12R4
M8S1	M4X62	M11	M14R4C2	M8C7	M15S10	M6C2	M8C4	M8C9
M15S10	M9C3	M15S14	M5C3	M8C9	M8C9	M8C9	M8C9	M8S5
M9C3	M15S6	M8C7	M10DR	M15S10	M9C3	M15S14	M8C2	M11
M10LP	M8S3	M4X62		M7I	M5.5C2		M11	M8C2
M15S11	M10DR	M8C13		M15S1	M4X3		M8S5	
	M8S4	M7GX		M15S13			M9C1	
	M10TB	M8C11					M8C14	
	M15S1							

Otherwise the model shows a lot of variables that are not very common among the other logistic regressions. This indicates that a different set of variables is important in age estimation when ages 23 through 35 are excluded. When age group 1 is included in the model, the phase of the auricular surface (M11) and the rib pit shape (M14R4C2) show up more frequently and appear to be important in differentiating out age group 1 individuals. At the other end of the age spectrum, when the oldest age group 6 is excluded from the data sample, the variables selected do not include M11, which again could indicate that the phase of the auricular surface's main purpose is to differentiate between the oldest and the youngest in the data sample. Other than these few patterns there is not much else that can be gleaned from the variables selected by the logistic regressions.

Two other things should be noted. First, besides the four variables that appear more frequently in these models, there are a lot of other variables that show up in one or two of the models. Thus there does not appear to be an overarching pattern to the variables selected in the logistic regression models beyond those highlighted in the table below. This suggests that different sets of variables may be more appropriate for different age ranges, and a method for age estimation should try to address this possibility. Secondly, the logistic regressions and decision trees outlined in Tables 5.9 and 5.10 lack a certain level of finesse for age estimation because they treat age as a categorical variable. They only measure success by the amount of observations that they classified correctly. However, as the age groups are defined by the author it is plausible that the particular age groups themselves may influence the variables selected for the logistic regressions. It is possible that the variables selected by these methods are different than the models discussed earlier simply because these models treat age as categories, while the earlier analyses treated age as a continuous variable.

Chapter 6: Discussion

The Use of Decision Trees

This research had its origin in two main facets. The investigation into age estimation methodology originated from a general frustration with age estimation methods in use when this research began. For years, anthropologists have been arguing about the difference between phase methods and component methods for age estimation. Articles in journals and papers presented at conferences discussed new ways of determining age, either based on new skeletal indicators or using a new mathematical method to get more reliable estimates (Bassed et al. 2011; Boyd et al. 2015; Cardoso and Rios 2011). Much of the work came from a similar dissatisfaction with the current standards in age estimation. The solution to getting more accurate and precise age estimates has always been to use multiple age indicators. More bones examined leads to more information collected, in turn leading to better age estimates. A holistic approach to age estimation—one that examines the complete skeleton instead of individual age-related markers—is believed to be the best way to get the best age estimates possible.

However, more data does not necessarily mean better results. It provides the potential, but more important than *more* information is information that is good and valuable to the issue at hand. Research testing the validity of established age estimation methods has found diverse results, ranging from good to not as good as required, anticipated, or wished (e. g., Lottering et al. 2013; Millán et al. 2013). This present research was built from a curiosity about the sub-stellar performance of established age estimation methods. However, rather than look at the performance of these methods on another validation sample, this research instead looks beyond the age determination methods and focuses on the variables used by the methods. The question

addressed is not how well the aging methods work, but rather how well the variables from those aging methods relate to age.

The idea of using decision trees to look at age indicator variables was inspired by an article written by Meindl and colleagues (1985). In this article, the authors described how the reader is supposed to determine into which phase a specimen should be classified. While the descriptions of the individual phases were given in short paragraph form, they also described a methodology for getting to the right phase. They described this as a step-by-step process (1985; 36-44); a process reminiscent of how a decision tree works. First you look at one particular part and, depending on what you find, determine the next step in the process. If certain conditions are met, you proceed one way with the evaluation, otherwise you go a different route.

This study used 94 different age indicator definitions gathered from 15 different methodologies—phase and component methods. While not technically holistic in nature (as only the four major age estimation sites were used and not the whole skeleton), this research should be exhaustive in terms of looking at the pubic symphysis, the auricular surface, the sternal rib ends, and cranial suture closure. After seeing what a wide array of age indicator variables used in conjunction with decision trees can derive, several interesting points of discussion come to mind. Possibly most important among these is why bother using decision trees for age estimation in the first place? This study does not present an age estimation method based on the use of decision trees, but it provides examples of what such a method might look like. Regression is one of the methods that can be used to combine multiple age indicators to get a single age estimate. Other methods include the transition method of Boldsen and colleagues (2002) and the older multifactorial method of Lovejoy and colleagues (1985a). Both of these methodologies are more complex than regression. In regression analysis, you use the data and get an equation. In order to

get an age estimation, you simply use that equation, plug in the data and calculate the age estimate. By comparison, the decision tree provides a flow chart, through which an observation can be traced to the terminal nodes where the age estimation is determined. When we compare these two methods, decision trees show some benefits and some drawbacks. A regression equation is very straightforward: plug the observed data into the equation, calculate, and you get the answer. A decision tree is not quite as straightforward, because you have to trace your observed data through the various splits in the tree. However, a decision tree is more intuitive and easier to understand than a regression equation, because all the decisions used to arrive at the age estimation are there in front of you and easy to follow. The easy to follow and intuitive nature of decision trees is also a major advantage over more complicated mathematical methods of age estimation.

Decision tree analysis also provides a benefit with respect to the actual process of how age estimates are made. Anthropologists have studied age indicators that change over time and have defined certain changes the indicator progresses through with age. However, it is impossible to know if everyone goes through the same skeletal changes. We know that human variation plays a role and that not everyone goes through the changes at the same speed. Some of these age-related changes may occur early on in life, possibly followed by long periods of static expression of the skeletal indicator, only to have it change again later on in life. A regression equation looks at the changes in the age indicator over the whole life span and considers aging a continuous process, progressing from one stage to the next unless otherwise specified. In such a way, an individual with a Todd phase 7 will always be aged older than an individual with a Todd phase 6, because the method defines that phase 7 comes after 6. Not only that, but the difference

between the two age estimations will always be a set discrete amount when only looking at the Todd phase.

By contrast, many of the assumptions made in a regression analysis are relaxed in the decision tree. As mentioned previously, decision trees have ways to handle missing data when they are constructed. If a tree is constructed with missingness treated as its own value, it is possible to estimate age of an individual with only some of the age indicators used in the decision tree. When you don't have a particular age indicator, you can bypass those splits in the tree and move on. So you could still get an age estimation if the pubic symphysis was not recovered or otherwise not observable. In regression analysis, any observation with missing data either needs to estimate that missing data by some means or eliminate that observation altogether. If the missing data cannot be estimated, the regression equation cannot be used.

A major benefit to decision trees is that they do not care about all the stages in an age indicator variable. This means that the decision tree does have to draw a line between all the different defined stages in an age indicator variable. For instance, the Gilbert and McKern (1973) component of pubic symphyseal rim formation defines six distinct stages of rim formation. When the decision tree looks at a variable like that it does not have to split the data into six distinct groups based on the state of their pubic symphyseal rim. Instead, the decision tree aims to split the observations into distinct groups based on where the stages actually differ with respect to age. After that split, the decision tree lumps together the observations on either side of the splitting indicator stage and treats them as the same in subsequent splitting. In our current example, if the tree found that the best split for symphyseal rim formation is between stages 3 and 4, then all the observations with rim stages of 0 to 3 would be lumped together in one group, and those with stages 4 and 5 would be lumped in a different group. The decision tree algorithm

would then attempt to split each one of those groups of observations further using a different discerning variable. It is even possible that the same variable could be selected again so that individuals with stages 4 and 5 could be separated. This would only happen however, if there was a significant difference in the observations based on that variable. While CRT trees always split data into two subsets along a single dividing value of the pertinent variable, CHAID trees can have more than one splitting value, but both methods perform a certain degree of lumping observations together even if they express different age indicator stages. The lumping of individuals observed to be different does suggest a loss of information inherent in the data. After all, if an age estimation method defines a different expression between stages 2 and 3, it does not seem right for the analysis to lump them back together. However, the important point to realize is that the observations are lumped together not because the method ignores their differences, but rather because it focuses on pertinent differences between the observations as it relates to the variable of interest. The difference in the expressions of stage 2 and 3 may not actually translate to a difference in the ages for members in those stages.

While it can be argued that the lumping together of different stages represents an ignoring of age-related information that discards important information, it can also be seen as a good thing. The lumping simplifies the age indicator assessment procedure for age estimation based on a decision tree because the assessor need only concern him/herself with the pertinent split of the age indicator. For the variable above, the investigator does not need to determine exactly to which symphyseal rim stage a specimen belongs. He/she only needs to determine if the specimen has advanced beyond a stage 3 or not. Furthermore, the lumping of different age indicator stages may help in better defining the age indicator, allowing for more consistent evaluation. If only a single—or a couple in the case of some CHAID trees—dividing value is found to be important,

researchers can focus their efforts on better defining what happens at that value and what characteristics separate the two adjoining stages. Such closer scrutiny could lead to better definitions of age indicator stages and reduce the amount of observer error that occurs in age assessment. The lumping can also be seen as a benefit, as all the differences perceived in the age indicator may not actually be statistically linkable to age and therefore only serve to muddle the assessment with unnecessary information.

In addition to not caring about every single defined age indicator stage, decision trees also do not evaluate all observations equally. If a regression equation is used, that equation is applied to every observation equally. Decision trees work differently. The only variable applied to every observation is the variable defined by the first split that divides the root node. After that, every other variable is only applied to a subset of the data. In the example above, the split between pubic symphyseal rim formation stages 3 and 4 is taken as the first split. This split creates two groups, one with symphyseal rims ranging from stages 0 to 3 and the other group having stages 4 and 5. We then take a second variable, such as rib end morphology, and find a significant split between stages 6 and 7 that applies to the stage 0 to 3 symphyseal rim group. During evaluation, the rib end morphology does not give two licks and a pennywhistle about any individuals with symphyseal rim stages of 4 or 5. More accurately, if observers see that their specimen has a rim stage 4, they do not have to evaluate the rib end morphology at all. They only need to worry about the variables that specifically pertain to their specimen and how they flow through the splits in the decision tree. In such a way a decision tree does not apply equally to all observations because some reach their terminal node much more quickly than others, and each observation only needs to be evaluated for the variables that are pertinent for its journey through the decision tree.

This characteristic of how decision trees function can also be seen as a double-edged sword. The decision tree does appear to ignore some information in the data by only applying *some* of the variables to *some* of the observations. However, the variables not utilized in this manner are not considered important enough to make a difference. Based on the state of age estimation, this researcher would argue that this characteristic of decision trees is more of an asset than a detriment. Most researchers interested in age estimation tend to weigh some age indicators more than others, relying more on those methods and indicators with which they feel comfortable. Combining multiple age indicators in an intuitive fashion naturally forces a researcher to weight the different age indicators against each other, leading to a single indicator being chosen as the "best" one. Using few or a single age indicator makes sense from a practical perspective as age estimation is often used in the field of archaeology where preservation is poor and archaeologists are not likely to have a whole skeleton to work with. In such cases age estimation methods that only rely on individual skeletal markers are much more useful, because it is rare to recover all skeletal elements that can be aged.

Nevertheless, in the larger picture of age estimation that is concerned with getting good age estimates based on multiple age indicators, decision trees provide an interesting potential benefit as well. Due to the way the trees split up the data and apply age indicators only to subsets, it is possible for a decision tree to find patterns between age indicator sites that are lost on other statistical methods. If we consider a particular variable that performs especially well on only a particular portion of the age continuum and poorly or only mediocre on the rest of the spectrum, such a variable as a whole would be considered average. Therefore, methods that look at the overall performance of variables to determine their validity would be unlikely to consider this variable. Decision trees may not. As they split the data, it is definitely possible for a decision

tree to select such a variable for its exceptional performance on a particular subset of cases. For example, if the auricular surface is only good at estimating ages of individuals over 40 years of age but provides no useful information for those who are younger, it does not sound very useful. However, if another variable is used in the decision tree that splits the data into individuals who are generally over 40 and under 40, then the auricular surface would become very useful as a secondary variable for the older portion of that data set.

So decision trees are capable of using the strengths of age indicator variables and ignoring their weaknesses exactly because they only apply these variables to portions of the data. In the simplest terms and as an example, it would be possible to have a decision tree that basically says: *Consider the sternal rib end morphology. If the ribs are young then look at the pubic symphysis for age estimation. If the ribs are mature, then look at the auricular surface for your age estimate, and if the ribs are very old you should use cranial sutures to determine the age of the individual.*

Because decision trees split up the data this way, they do not look for all the good variables that relate to age. Instead, a decision tree selects a set of variables to complement each other. If a variable under consideration relates strongly with the target variable but does not offer any more information than is covered by the variables already in the tree, that variable is not going to be selected. It will likely be passed over for a variable that is less strongly related to the target variable, but has a stronger impact on the performance of the decision tree. This process prevents the decision tree from selecting very strong variables that have the same relationship to age. In this research, many of the 94 variables utilized pertained to the same age indicator. For instance the development of the ventral bevel of the pubic symphysis is defined multiple times (Boldsen et al. 2002; Chen et al. 2008; Gilbert and McKern 1973; Hanihara 1978; McKern and

Stewart 1953). Each one of these researchers described the changes to the ventral bevel differently, but overall the changes still pertain to the same skeletal age indicator and therefore the same age-related changes. A decision tree would be unlikely to pick such a variable multiple times because they are all likely very similar to each other. Instead, the decision tree would look to other variables that would complement the development of the ventral bevel. Any other statistical method would also be unlikely to choose more than one of these variables.

When you look at a decision tree it is possible to consider the variables at the top of the tree as more important, which they are in a way because they have influence over a larger number of observations. These variables split the data into large general age groups and as such could be considered "gross age estimating variables." On the other hand, the variables toward the bottom of the tree and the ends of the branches only have a say over a small proportion of the data set. These variables split the large general age groups into more defined specific groups and as such could be considered "fine tuning variables." In this way, every decision tree selects a set of variables that are both gross and fine tuning age estimating variables; it is the teamwork of these variables that gives the model its power.

The variable selection process of the decision tree algorithms can be confusing as well. Since the decision tree only selects variables based on the data subset in a particular node, if that subset is changed then any subsequent variables would also be changed. Similarly, if a variable is substituted into the decision tree at a particular point for a different one, the new variable will split the data differently, therefore creating two different child nodes. These child nodes in turn would be split by yet new and different variables, and so on. Thus, substituting a single variable for another will invariably change the tree both in shape and the variable it chooses as the best with respect to predicting power of the target variable. This drastic transformation can be seen by

comparing the CRT 10/5 decision tree (see Figure 5.14) with the CRT 10/5 decision tree without observation 15 (see Figure C.5). The two trees use a different initial variable which create different child nodes, and so on. This avalanching effect on decision trees makes it very difficult to test the effect a single variable has on the model. Since changing a single variable likely changes a majority of the tree it becomes very difficult to ferret out the practical effect of any single variable on the age estimation process.

In this author's opinion, the benefit in the use of decision trees for age estimation is that they have to make very few assumptions about how the age indicators actually relate to age. They also show promise in their ability to sift through a large amount of age indicators and reduce the amount of data required. They are able to highlight patterns between age indicators by selecting complementary variables, and also highlight patterns within a single age indicator by showing when the most age-related changes to the indicator occur. They can also be used to test various subsets of age indicators to find the pertinent pattern between indicators the researcher is interested in. They can test the relationship between the indicators of a single component age estimation method, or just the age indicators they consider worthwhile. Their simple and easy to follow diagrams allows for an easier understanding of the patterns present in the data.

The Variables

This study utilizes 94 age indicator variables collected for 176 individuals of known ages. That is a very large number of variables compared to the number of observations. The large number of variables, although collected with the best of intentions, actually proved to be detrimental during the analysis. As mentioned previously, several of the variables were derivations of each other, basically outlining a different way to classify the same thing.

However, they are still looking at the same site on the skeleton and describing the age-related changes perceived there. As such, many of the variables used in this study are highly correlated to each other, meaning they are very similar in the data they provide and would be acceptable substitutes for each other.

The high degree of intercorrelation between the age indicator variables means that many of the variables chosen to be used by the decision tree algorithm probably just slightly outperformed another variable that did almost as well. It is therefore perfectly reasonable to argue that there are some variables in the data set that did not show up in any or many of the decision trees, yet are perfectly adequate for age estimation. The author is not claiming that all the variables not selected in the decision trees are worthless. They just did not offer more information than the other variables already in the model.

Even if good age-related variables were not selected to be part of the decision trees because they were overshadowed by other, potentially slightly better, variables, the end results of this study still hold. Variables that showed up more frequently should be considered to contain valuable information for the age estimation process.

Age Distributions

The sample utilized in this study consisted of 176 individuals, selected to obtain approximately 30 specimens for each 5 year interval. Figure 6.1 shows the histogram of the sample illustrating how top heavy it is. Fortunately for society, yet unfortunately for researchers of age estimation, the Bass Donated Collection currently contains only a limited number of adults under the age of 30. The data set in this study only contains 12 individuals under the age

of 30, and just 16 between the ages of 31 and 35. The youngest is 23 years of age and the oldest is 60.

This sample distribution likely has some influence over the results presented here. Some of the age indicator variables used in this research come from age estimation methods that were established from a relatively young data set. For instance, the analysis of McKern and Stewart (1957) is frequently criticized because it was established from a sample of young soldiers, with the oldest individual in their sample being of age 50, and only 2 individuals over the age of 42. The youngest individuals in their data set are 17 years old. Likewise, the data used to establish the definitions outlined by Hanihara and Suzuki (1978) is based on individuals that range between the ages of 18 and 38. Compared to those data sets, the data set utilized in this research

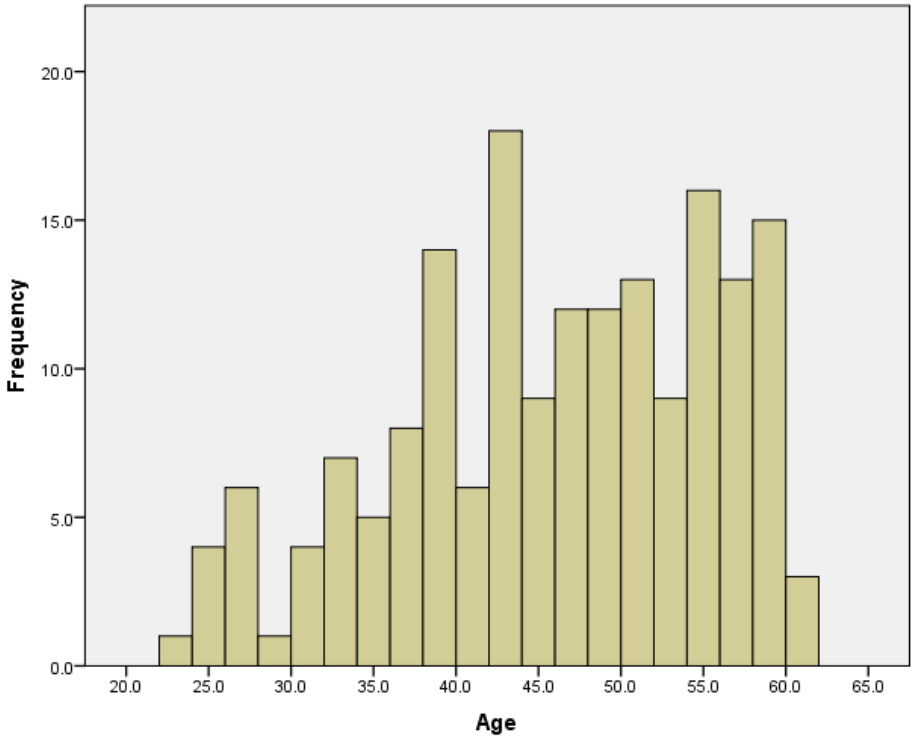


Figure 6.1: Histogram of the age distribution of the study sample

is different in its distribution because most individuals are actually over the age of 35. Therefore, when the decision tree (or any other statistical method) is looking for significant variables as they relate to the age distribution used in this study, variables previously determined to perform better on older individuals will be chosen as more significant simply because there are more older people in the data set. A variable that can factor out the 12 individuals in their 20s will be less valuable to the method than a variable that would specifically group the thirty individuals from the data set between the ages of 55 and 60. Methods that were built on data sets of young individuals (e.g., Hanihara and Suzuki 1978; McKern and Stewart 1957) will define variables that are much more specific to differentiating between individuals from their own data set. Variables from such methods are unlikely to be chosen as the main age estimation variables for estimating age in data sets of older individuals, such as the one used in this study. For this reason, it is unlikely for the major variables in the decision tree of this study to be based on a variable established on a young data set.

A similar argument can be made about the aging sites used and why the auricular surface and the rib end morphology showed up more frequently in the decision trees than variables based on the pubic symphysis. Besides methods being established based on different age distributions, the specific age indicator should also be considered. Age-related changes to the pubic symphysis have been studied and described extensively (e.g., Todd 1920). These changes happen in the early years of maturity and these early changes are distinguishable from each other by major morphological changes. However, after the dorsal plateau is completed around the mid to late 30s (Brooks and Suchey 1990), the observable changes to the pubic symphysis become much less distinct and the descriptions of these changes become more vague and subject to interpretability. The auricular surface as well as the rib end morphology suffer from a similar

generalization toward the end of the recorded age-related changes. However, for these age indicators the age changes that are clear to define persist longer than for the pubic symphysis. The pubic symphysis is more specific with regard to changes that happen in the 20s and 30s, and the auricular surface and rib morphology descriptions exhibit distinguishable features into the 40s. After their ages of posterity, all methods show a decrease in specificity and accuracy with regard to age estimation. The focus of age-related changes in younger individuals is a potential reason why the pubic symphysis was not selected more frequently for use in the decision trees. Because the data set has many more older individuals, well documented age-related changes in younger individuals seen in the pubic symphysis are not as valuable. By comparison, the auricular surface and the rib morphological changes are documented more specifically into later portions of age and as such are more valuable to a data set that has an older age distribution.

Age Indicator Variables

One of the major interests in this study was to determine if there were any good variables out there and which ones they were. The term "good" here is subjective. For most age estimation enthusiasts, a "great" age indicator would be one that relates strongly with age and one in which the different stages expressed by the age indicator have a narrow and consistent age during which changes are clearly observable. Also, it should have stages that range from age 20 to age 80. All that should lead to age estimations with ranges of 5 years—obviously the exact age would be preferred, but let's not get overly optimistic. If that is the case, then a "good" age indicator should have *some* of these characteristics *some* of the time with *some* of its stages being *some*what applicable to *some* ages. While this researcher agrees with this description of a "good" age indicator, he also believes that there is something frequently overlooked when age indicators

are established. That is the consistent repeatable application of the age indicator. Science is about experiments with repeatable verifiable results. Therefore, a scientific age estimation method should use age indicators that can be applied by various scientists to the same skeletal element and generate the same age estimate, or at the very least the same age indicator stage. Interobserver errors are well documented with respect to age estimation (Baccino et al. 1999; Berg 2008; Buckberry and Chamberlain 2002; Saunders et al. 1992). A good age indicator would want to minimize the amount of interobserver error, if not outright eliminate it.

We have already mentioned that some of the age indicator variables used in this study measure the same thing. The development and formation of the ventral bevel and rampart on the pubic symphysis is a good example. No fewer than four different methods used in this research contain a variable that describes the stages of the ventral bevel. Hanihara and Suzuki (1978) describes each stage in five words or less. McKern and Stewart (1957) use single, concise sentences to describe each of their stages. Chen and colleagues (2008) use a sentence fragment. Boldsen and colleagues (2002) use long sentences and, in part, extensive paragraphs to describe the stages they defined in the ventral aspect of the pubic symphyseal margin. All of these researchers looked at the same age indicator site and presumably observed the same age-related changes, yet they chose a different way to describe them to the reader. Is the description given in a long paragraph easier to understand than a description given in a few words? Possibly. Does a concise description make it easier to evaluate a pubic ventral margin than a paragraph? Maybe. Which one of these four ways of classifying the ventral pubic development is the best? It is not a competition and they are all trying to do the same thing. However, it is easy to see how these different methods with their different definitions of the same age-related changes could lead to different evaluations of pubic symphyseal ventral bevel and rampart formation. There may not

even *be* a best way to describe age-related changes to the skeleton, but a "good" age indicator needs to put forth definitions that not only describe features that can be identified, but can also be understood and applied by the person using the definition to identify the features described.

In this study, the researcher used definitions put forth in 15 different age estimation methods on four major age estimation sites. The methods ranged from using short to extensive descriptions about the various age indicator variable stages. Using the information presented in the descriptions and accompanying articles, the author collected the data on 94 differently defined age indicator variables. Variables with unclear definitions were recorded as interpreted by the researcher, based on what he believed the authors of the methods meant to describe. The author of this study is not above observer error, and as such did not presume he would record all variables correctly. However, care was given to collect the data as consistently as possible. Due to the potential for observer error and misinterpretation of some age indicator stage definitions, this study did not attempt to validate any of the age estimation methods used in this study.

This study identified age indicator variables that related strongly with chronological age. The observer, using the definitions provided, actually identified stages in the skeletal sample that were related to age. In other words, the top variables found in this study have three characteristics: 1.) they are correlated with age; 2.) they define stages that distinguish between age groups; and 3.) the observer, using the definitions, identifies significant age differences.

Jackknife Decision Trees and Logistic Regressions

The various decision trees and logistic regressions that were run with jackknifed subsamples give insight into the roles different variables play in the age estimation. The first major insight comes from the variables that appear again and again. In particular the phase of the

auricular surface (M11) is one of the most common variables that shows up time and again in both decision trees and logistic regressions. It is one of the few variables in the data set that has more than 5 categories. Having a larger number of categories naturally means that the variable has a larger number of potential splitting points that can be found to be significant by the decision tree growing algorithms. In both regression and logistic regression the larger number of categories of this variable means that that single variable contains more information than a variable with less categories, such as a binary one. Obviously the variable still needs to contain valuable information with respect to the dependent variable in order to be utilized in the models. However, just by having more categories the variable increases its chance that one of its categories will be important for age estimation and therefore the whole variable will be selected. For instance in the CHAID 30/10 decision tree with all observations, but also in the CHAID tree without age group 3 or 4, the variable M11 is selected. Yet in all three trees the splitting points are the same. Stages 1, 2, and 3 are grouped together as are stages 4, 5, and 6, and stages 7 and 8 are grouped into a third category. While the decision tree did not find it necessary to differentiate between all the categories of the auricular surface phase, some of its categories provide useful information to the tree.

All four of the jackknife decision trees at either end of the age spectrum (i.e., those excluding the oldest and the youngest age groups) share a few characteristics. They show lower mean absolute error values for the training sample and very large error value for the jackknifed observations. As already mentioned, the small MAE value for the training sample results from the reduced variation in the data, due to the elimination of a chunk of observations at one end of the age range. Besides showing that the decision tree is unable to estimate ages accurately for the jackknife group, the large jackknife values also indicate that decision trees are not good at

extrapolating age estimates outside the range of their training sample. This fact needs to be remembered when creating a decision tree model or using a decision tree to evaluate a skeleton. A decision tree model needs to specify for which individuals it is useful. Other statistical methodologies may prove better at extrapolating age estimates outside the range from which they were constructed. However, due to the non-linear nature of the age estimation indicators, as well as the complexity of the aging process, most (if not all) methods may find it extremely challenging to predict what happens outside the methods' range of training.

The four decision trees with the end age groups removed also are quite distinct from the overall decision tree utilizing the complete data set as well as the decision trees lying between them (i.e., those excluding age groups 2, 3, 4, and 5). These trees show that different variables are good for age estimation for different subset of the data, in particular when portions of the data are removed from the ends of the age spectrum. One particular characteristic these trees show is that none of them utilize the most commonly used variable in this data set--the phase of the auricular surface (M11). This is also true for the two jackknife logistic regressions excluding age group 1 and 6. While all four of the intermediate logistic jackknife regressions use M11, neither the regression excluding age group 1 nor age group 6 use this variable in their model. The same is true for the four end-spectrum decision trees. Several of the mid-spectrum decision trees do use M11 as part of the tree. These characteristics imply two things. First, since M11 only shows up when the whole age range is being assessed by a model, the variable appears to be useful for large scale differentiation between major age groups. In this data set apparently you need to be looking at the age range from ages 23 to 60 in order for M11 to be considered a good differentiating variable. If the age range is reduced to 35 to 60 years or 23 to 55 years, a different

variable does a better job at defining the major differences between the old and the young individuals.

Due to the small number of young individuals in this data sample, age group 1 ended up being an almost 15 year range including individuals from ages 23 to 35. Compared to the whole age range the sample drawn from age group 1 represents almost half of the age range from 23 to 60. Hence, the jackknife models in this study are not informative about the differences in important age estimation variables pertaining to the younger half of the studied age range, since most of these were treated as one group. Due to the limited sample size, the author did not look for differences between individuals in their 20s and those in their early 30s. This is obviously an area that needs to be investigated further, but the data in the current sample is insufficient for this purpose. If the sample had enough data to create 5-year interval jackknife groups for the young half of the age range (i.e., 20-25, 25-30, and 30-35), analyses would likely find different results for these groups of individuals. The findings in this research suggest that for these groups variables of the pubic symphysis and the rib morphology would likely supplant auricular surface variables as those of primary importance. Such findings could lead to the consideration of developing different age estimation methods for different portions of the age range, such as those under 40 years and those over 40.

Furthermore, the uniqueness of the models based on the reduced age range samples (35-60 and 23-55), indicate that different variables are important for different ages. Several different patterns appear in the various jackknife decision trees.Variables pertaining to the pubic symphysis are only useful for younger groups of individuals. The auricular surface (especially M11) and the sternal rib end morphology (particularly pit shape (M14R4C2)) are generally used for the initial splits in the decision trees. Thus, these variables and age indicators are useful for

large scale differentiation between young and old individuals. Cranial suture closure variables and other auricular surface variables are generally used to define age groups for the older portion of the sample, those between ages 40 and 60.

The repeated use of the phase of the auricular surface variable (M11) in the middle age-spectrum jackknife decision trees brings up a potential problem for identifying useful age indicator variables. Since certain variables appear to be frequently selected for defining the young group of individuals in this data sample, it is possible that by including the young individuals in the sample, a variable like M11 will be selected to factor out the young individuals. However, by doing so that variable may actually overshadow a different variable that would be more useful for age estimation if the young individuals were not in the data set. Thus, the variables selected by the jackknife models that exclude age group 1 are obviously more important for the older individuals, but also represent a unique set of variables that are not obscured by the need to include a variable like M11 to simply separate the data into younger and older individuals. It is therefore possible to argue that the inclusion of the younger individuals in these analyses, while making them applicable to a wider age range, actually manages to obscure useful variable relationships.

It is also interesting to note that few of the jackknife models show the same pattern. The CRT trees with the age groups 3, 4, and 5 held out are very similar. But for the most part, most of the models, including the logistic regressions do not show consistent variable selection. The patterns that do repeat are general patterns about the type of age indicators that are used for particular situations, rather than particular age indicators: auricular surface variables for older individuals, rib and pubic symphysis variables for younger individuals, and cranial sutures for specific smaller groups of individuals. Also there are certain common variables that show up

repeatedly in different models, generally serving the same function. However, by and large different models select different variables. This makes it hard to distinguish any realistic pattern in the variables selected. The lack of consistently selected variables as well as the prevalence of the auricular phase variable (M11) to separate young from old individuals suggests that a different modeling approach may be more useful for age estimation. A combination of different methods may prove more successful at modeling the age progress, than a single decision tree or even a single multiple regression. A more optimal age estimation method may be a model where one uses the auricular surface to determine if you are working with an older or younger individual. Based on that decision, a different set of variables may prove more useful for each of the sub-ranges of the age continuum. Alternatively, it could also be possible that ages for the young and old individuals are modelled by different statistical processes altogether. However, such a two-step age estimation modeling approach is something that will need to be tested in a different study.

Specific Results

This research used 94 different variables based on four sites in the skeleton and looked at how they related to the age of 176 specimens. Some of these variables pertain to the same features and, as such, the data does not actually measure 94 different features. Further, all 94 of these variables were defined in such a way as to record some feature that changed during the course of an individual's life. With so many variables all aimed at outlining one age-related change or other, at least some of these variables should be found to be important for age estimation. It might even be reasonable to expect that a model using these variables to estimate age would be pretty good, since there is such a large amount of information provided, yet even

with all these data and all these variables, the decision tree models constructed in this study are mediocre at best. Most of the models—both decision trees and regressions—showed average estimation errors around 5 to 6 years and only a few of the models accounted for more than 60% of the variation in the data (see Table 5.4). Furthermore, the maximum amounts of error were around 18 years of over- and underestimation of age. Even a 13 variable regression equation only accounted for 60% of the variance in age and had an average aging error of 4.7 years.

So why are the results not stellar? Some of the issues have already been discussed in previous sections. The fault may lie in observer error during data collection. Decision trees may not be the right method for skeletal age estimation. The data may not be large or balanced enough to allow regression or decision trees to find good reliable patterns. It could also be possible that this level of accuracy is as good as age estimation can be using the variables included in this study. Ninety-four variables seem like they would be enough to exhaust the amount of data that can be collected from the four aging sites used here. If this is the best that the combination of the pubic symphysis, the auricular surface, the sternal rib end morphology, and cranial suture closure provides, then in order to improve age estimation further, other age-related skeletal indicators need to be identified that add to the information of age changes in the skeleton.

Several decision trees presented in this paper show unexpected patterns. For instance, CHAID 30/10 (Figure 5.4) shows an interesting split under node 3. For the split it uses the texture of the auricular surface scores defined by Buckberry and Chamberlain (2002). The interesting part is that the higher scores actually estimate a lower age. Individuals with a score of 4 or less are estimated to be 55 years old while individuals with scores of more than 4 are only estimated to be 50. The stages outlined in the method are arranged in supposed chronological

order of progression. Therefore, the pattern seen in the decision tree makes no sense and is likely caused by observer error. In the method definitions, stages 4 and 5 are characterized by "dense bone" while earlier stages show no dense bone. Thus, the observer likely misidentified what dense bone is supposed to look like during data collection.

This example highlights what the decision tree accomplished in this study. The variable is selected in a few decision trees. Even though the observer appeared not to have applied the definitions as the authors intended, the definitions of the age indicator nevertheless applied to specimens consistently enough and showed a difference between individuals with "dense" bone and those without. This instance highlights to the observer that he misinterpreted something in the age method, and it tells the authors of the original methodology that there could be an issue with consistent application of their method. On a positive note, the decision tree also indicates that "dense bone" is a characteristic of the auricular surface with strong ties to age.

A second example of decision trees highlighting a disparity between the observer data and the defined method is shown in the third split of the phase-only CHAID 30/10 decision tree (see Figure 5.19). With regard to the phases of the pubic symphysis as defined by Todd (1920), the decision tree shows that 27 individuals with phases of 6 or less have a higher average age than 23 individuals with phases 7 and 8. The definitions by Todd (1920) are more complex than those by Buckberry and Chamberlain (2002) and do not present an obvious reason why this pattern occurred in the data. However, the decision tree highlights an area in an age estimation method that should be considered more closely in order to determine what the issue might be.

The other results of this study are the variables identified as being the best for use in age estimation. Table 6.1 is an abridged version of the results table 5.5. This paper argues that even if the decision trees do not have good performance with respect to age estimation, the variables

Table 6.1: Eight most common variables used in distinct decision trees and regression models.

Code	Variable	Bone	Type	Counts
M11	Auricular surface	Auricular	Phase	6
M14 R4 C2	Pit shape	Ribs	Component	6
M9 C4	Degree of Macroporosity	Auricular	Component	5
M15 S1	Midlambdoid closure	Sutures	Component	5
M4 X4	Dorsal Margin	Pubis	Component	4
M12.5 R4	Sex-specific rib morphology	Ribs	Phase	4
M15 S6	Midcoronal closure	Sutures	Component	4
M15 S8	Sphenofrontal closure	Sutures	Component	4

selected by the decision trees should still be considered better than those variables that did not make it into the decision trees. Due to the sample's age distribution, these variables should be considered to be especially useful in age estimations for individuals of advanced age. The two top variables that made appearances in the various decision trees were the phase of the auricular surface and the pit shape at the sternal ends of the fourth rib. We already discussed possible reasons why these two variables could have been chosen more frequently than any other. Both the changes in the auricular surface and the rib morphology have been defined for a larger age range than the pubic symphysis. This fact, together with the advanced age of the data, gives one explanation as to the prolificity of the top two variables. Pubic symphysis variables do not show up as frequently in the decision tree models because they are focused on age changes that occur in only a small portion of the data set. The one pubic symphysis variable in the top 8 is the development of the dorsal aspect of the symphysis. This aspect is one of the last distinctly defined changes the pubic symphysis undergoes and is therefore more applicable to an older data set than variables that describe earlier changes in the pubic symphysis. Three cranial suture variables also appear in the top 8. The three sutures are located at different locations on the cranium, with the midlambdoid at the back of the skull, the midcoronal on the top, and the

sphenofrontal on the sides. These three sutures are also quite different in structure. The coronal suture is very simple, straight and wide. By comparison, the lamboidal suture is tighter and often very complex. The sphenofrontal suture is completely different because it is very small, fine, and very tight. The differences between these sutures explains why all three appeared so frequently in the models. Their different locations and the different types of sutures likely relate to different times or events that cause these sutures to fuse. Furthermore, when looking at the location of these variables in the decision trees we find that they are located near the ends of branches. As such, they are filling a very specific role for a small proportion of observations.

Readers of this study may wonder why their most commonly used aging variables may not be on the list, even if the variables provide them with consistently good age estimates. Some of the reasons for this have already been discussed. The variable may not have been as applicable to this data set as other variables. It might fulfill the same purpose as another variable, but does it just slightly worse from a mathematical standpoint. It is also important to remember that a single variable might be good at age estimation by itself, but a set of variables to complement each other will generally perform better. It is quite possible that two mediocre age estimation variables complement each other so well that they outperform a single good age estimation variable. All in all, as seen in Table 6.2, 37 out of the 94 variables that were available made at least one appearance in one model or another. However, the variables only chosen once or twice were likely chosen because of specific circumstances that allowed them to shine. Such variables are important to include in models because they may provide information overlooked or not addressed by the main variables. Due to the very minor role these variables take in the overall age estimation process, however, it is hard to argue that any one of them is better than another. It

Table 6.2: Breakdown of variable occurrence by aging site.

Age Indicator Sites	Number of Variables	Variables in Models	Number of Occurrences	Average Number of Times used
Auricular Variables	30	14 (47%)	32	2.3
Rib Variables	6	4 (67%)	14	3.5
Pubic Symphysis Variables	39	11 (28%)	19	1.7
Suture Variables	19	8 (42%)	19	2.4
Total	94	37 (39%)	84	

all comes down to the data set or the individual for whom age needs to be estimated. Overall, the most common variables selected should still be considered the most useful for an overarching age estimation method, especially for advanced ages.

Future Directions

There are a variety of directions that research in the field of age estimation can take. Validation of published methods is one route, but only adds a little information to the field. A major area of inquiry would be the source of inter- and intraobserver error. In addition to recording how frequently the observer error occurs, it would be quite interesting and potentially more promising, to investigate why it occurs and where it originates. Finding these answers may lead to a way to reduce the amount of observer error, which in turn should decrease the inconsistency of age estimation methods. Decision trees may prove helpful in this aspect. This research has only scratched the surface of the use of decision trees for age estimation.

Future research should include a larger proportion of younger individuals to test if the pubic symphysis would prove more useful in a data set with a more even age distribution. An increase in sample size, especially in the younger ages would also allow testing of narrower

jackknife samples in those under the age of 35. Taking out smaller jackknife groups will allow for closer examination of the age indicator variables that are important for those age ranges. It was already noted that the logistic regressions in this study did not impress and that the consideration of interaction variables could improve the logistic results. A larger data sample will allow for testing of more interactions. Nevertheless, due to the large number of possible interactions, it may be necessary to pick individual interactions to test rather than all of them. Picking these useful interaction variables may come down to expertise on the part of the anthropologist. Alternatively, looking at the results of various decision trees may also suggest certain interactions to test based on the variables and splitting criteria used in the decision trees. An increase in the number of female individuals in this data sample would also allow investigation of sex differences in age estimation and aging variables. Some age estimation methods have been established for use with a particular sex, but other methods claim to be applicable to both. Testing the differences between the sexes can be done individually, but more telling would be a jackknife approach where one of the sexes is treated as the validation sample for the method developed on the other sex. This methodology could shed light on if differences in aging exist between the two sexes and if different age estimation variables are more helpful in determining the age of one particular sex.

An alternative approach similar to the jackknifing used in this study would be to study how the variables selected for age estimation change as one expands the age range under investigation. Starting with an analysis of individuals between 55 and 60 and expanding the age range in successive analyses could show at what point in the age range different variables become important. The changes in variables as the age range tested changes would reveal what

variables are important for which age ranges and also what function a particular variable serves in the model or decision tree.

Furthermore, decision trees can be used to find complementary sets of variables. Such sets of variables should be validated and might lead to new holistic approaches to age estimation. They can also be used to hypothesize interaction variables to be tested by logistic regression or another method. Decision trees could also be used to determine exactly between which stages of an age indicator the most important differences lie. These stages should then be redefined so that the critical points in differentiation are made clear and are easy to replicate. As illustrated in this study, decision trees can also be used to find instances of observer error. By showing where inconsistencies occur between the age indicator definitions and the age indicator data collection, it would be possible to strengthen age estimation methods. Decision trees are very well equipped to sift through a large amounts of data and variables and present patterns found in an easily observable and interpretable manner.

Something more complex to consider for age estimation would be a combination of methods. Already mentioned earlier during the discussion of the inconsistent jackknife results, a rather more complex approach to age estimation would be a sort of decision tree of age estimation methods. In this scenario one method would be used to study age, but rather than estimate age the results would point you to one of a set of secondary age estimation methods, depending on the results of the meta-method. This approach could allow for different methods to be applied to different age ranges and the results in this study already suggest that different variables work better on different age ranges. While complicated, such a two-step approach to age estimation may improve current age estimation methods.

Chapter 7: Concluding Remarks

Decision trees offer interesting possibilities to the realm of skeletal age estimation. As anthropologists attempt to increase the reliability of age estimation, more and more advanced aging methods will be introduced. These advanced methods will need to make certain assumptions about the age variables and their relationship to chronological age. The fact that accurate age estimates are so hard to obtain should tell us that the process of age estimation is not as straightforward as we like to think it is. Therefore, we need to take a step back to consider what the cause of our problems may be. Decision trees allow relaxation of certain assumptions that we routinely make about age estimation. Variables measuring the aging progress do not have to be treated as continuous. We do not have to assume that all stages follow each other sequentially, that they are all distinct, or that each stage provides important information the other stages do not. More importantly, decision trees offer a way to test age estimation methods and present results in easy to understand diagrams.

This research showed that:

--for individuals between the ages of 35 and 60, the pubic symphysis is not as helpful as other major age estimation sites such as the auricular surface and the sternal rib ends.

--the auricular surface phase method and the development of the sternal pit shape of the fourth rib are strongly linked to chronological age and should likely be included in any holistic approaches to age estimation, or any aging method developed for older individuals.

--the best aspect of the pubic symphysis to focus on in older individuals is the dorsal aspect.

--certain cranial suture closure sites are more helpful than others and provide information for age estimation not covered by postcranial aging sites.

--different age indicators are important for age estimation of different ages.

Lastly, we have to realize that bad age estimation methods cause a lot more problems than just bad age estimates. Age estimation methods are used in a variety of research questions and are therefore a cornerstone to many investigations. For those investigations to be reliable and to come up with trustworthy conclusions, the age estimations have to be reliable as well.

References Cited

Acsádi GY, and Nemeskéri J. 1970. History of Human Life Span and Mortality. Budapest: Akadémiai Kiadó.

Ashley-Montagu MF. 1938. Anthropometry: Aging of the Skull. American Journal of Physical Anthropology 23(3):355-375.

Aykroyd RG, Lucy D, Pollard AM, and Solheim T. 1997. Technical Note: Regression Analysis in Adult Age Estimation. American Journal of Physical Anthropology 104:259-265.

Baccino E, Ubelaker DH, Hayek LA, and Zerilli A. 1999. Evaluation of Seven Methods of Estimating Age at Death from Mature Human Skeletal Remains. Journal of Forensic Sciences 44:931-936.

Bassed RB, Drummer OH, Briggs C, and Valenzuela A. 2011. Age Estimation and the Medial Clavicular Epiphysis: Analysis of the Age of Majority in an Australian Populatin using Computed Tomography. Forensic Science, Medicine, and Pathology 7:148-154.

Bedford ME, Russell KF, Lovejoy CO, Meindl RS, Simpson SW, and Stuart-Macadam PL. 1993. Test of the Multifactorial Aging Method Using Skeletons With Known Ages-at-Death From the Grant Collection. American Journal of Physical Anthropology 91:287-297.

Berg GE. 2008. Pubic Bone Age Estimation in Adult Women. Journal of Forensic Sciences 53(3):569-577.

Bocquet-Appel J-P, and Masset C. 1982. Farewell to Palaeodemography. Journal of Human Evolution 11:321-333.

Boldsen JL, Milner GR, Konigsberg LW, and Wood JW. 2002. Transition Analysis: A New Method for Estimating Age from Skeletons. In: Hoppa RD, and Vaupel JW, editors. Paleodemography Age Distributions from Skeletal Samples. Cambridge: Cambridge University Press. p 73-106.

Boyd KL, Villa C, and Lynnerup N. 2015. The Use of CT Scans in Estimating Age at Death by Examining the Extent of Ectocranial Suture Closure. Journal of Forensic Sciences 60(2):363-369.

Brooks S, and Suchey JM. 1990. Skeletal Age Determination Based on the Os Pubis: a Comparison of the Acsádi-Nemeskéri and Suchey-Brooks Methods. Human Evolution 5:227-238.

Brooks ST. 1955. Skeletal Age At Death: The Reliability of Cranial and Pubic Age Indicators. American Journal of Physical Anthropology 13:567-597.

Buckberry JL, and Chamberlain AT. 2002. Age Estimation from the Auricular Surface of the Ilium: A Revised Method. American Journal of Physical Anthropology 119:231-239.

Cardoso HFV, and Rios L. 2011. Age Estimation From Stages of Epiphyseal Union in the Presacral Vertebrae. American Journal of Physical Anthropology 144:238-247.

Chen X, Zhang Z, and Tao L. 2008. Determination of male age at death in Chinese Han population: Using quantitative variables statistical analysis from pubic bones. Forensic Science International 175:36-43.

Dudar JC, Pfeiffer S, and Saunders SR. 1993. Evaluation of Morphological and Histological Adult Skeletal Age-at-Death Estimation Techniques Using Ribs. Journal of Forensic Sciences 38:677-685.

Dwight T. 1890. The Closure of the Cranial Sutures as a Sign of Age. Boston Med and Surg J 122:389-392.

Falys CG, Schutkowski H, and Weston DA. 2006. Auricular surface aging: worse than expected? A test of the revised method on a documented historic skeletal assemblage. American Journal of Physical Anthropology 130:508-513.

Forensic Anthropology Center. 2015. WM Bass Donated Skeletal Collection. University of Tennessee, Knoxville. http://fac.utk.edu/collection.html

Galera V, Ubelaker DH, and Hayek LA. 1998. Comparison of Macroscopic Cranial Methods of Age Estimation Applied to Skeletons from the Terry Collection. Journal of Forensic Sciences 43(5):933-939.

Galera V, Ubelaker DH, and Hayek LC. 1995. Interobserver Error in Macroscopic Methods of Estimating Age at Death from the Human Skeleton. International Journal of Anthropology 10:229-239.

Garn SM, Nagy JM, Sandusky ST, and Trowbridge F. 1973a. Economic Impact on Tooth Emergence. American Journal of Physical Anthropology 39:233-238.

Garn SM, Sandusky ST, Rosen NN, and Trowbridge F. 1973b. Economic Impact on Postnatal Ossification. American Journal of Physical Anthropology 38:1-3.

Gilbert BM, and McKern TW. 1973. A Method For Ageing the Female Os Pubis. American Journal of Physical Anthropology 38:31-38.

Godde K, and Hens SM. 2012. Age-at-Death Estimatin in an Italian Historical Sample: A Test of the Suchey-Brooks and Transition Analysis Methods. American Journal of Physical Anthropology 149:259-265.

Gowland RL, and Chamberlain AT. 2002. A Bayesian Approach to Ageing Perinatal Skeletal Material from Archaeoogical Sites: Implications for Infanticide in Roman Britain. Journal of Archaeological Science 29:677-685.

Haglund WD. 1997a. Dogs and Coyotes: Postmortem Involvement with Human Remains. In: Haglund WD, and Sorg MH, editors. Forensic Taphonomy: The Postmortem Fate of Human Remains. Boca Raton: CRC Press. p 367-381.

Haglund WD. 1997b. Rodents and Human Remains. In: Haglund WD, and Sorg MH, editors. Forensic Taphonomy: The Postmortem Fate of Human Remains. Boca Raton: CRC Press. p 405-414.

Haglund WD, Reay DT, and Swindler DR. 1988. Tooth Mark Artifacts and Survival of bones in Animal Scavenged Human Skeletons. Journal of Forensic Sciences 33(4):985-997.

Haglund WD, Reay DT, and Swindler DR. 1989. Canid Scavenging/Disarticulation Sequence of Human Remains in the Pacific Northwest. Journal of Forensic Sciences 34(3):587-606.

Hanihara K. 1952. Age Changes in the Male Japanese Pubic Bone. Journal of Anthropological Society of Nippon 62(698):245-260.

Hanihara K, and Suzuki T. 1978. Estimation of Age from the Pubic Symphysis by Means of Multiple Regression Analysis. American Journal of Physical Anthropology 48:233-240.

Helfman PM, and Bada JL. 1976. Aspartic acid racemization in dentine as a measure of ageing. Nature 262:279-281.

Hens SM, Rastelli E, and Belcastro G. 2008. Age Estimation form the Human Os Coxa: A Test on a Documented Italian Collection. Journal of Forensic Sciences 53(5):1040-1043.

Holman DJ, Wood JW, and O'Connor KA. 2002. Estimating Age-at-Death Distributions from Skeletal Samples: Multivariate Latent-Trait Approach. In: Hoppa RD, and Vaupel JW, editors. Paleodemography Age Distributions from Skeletal Samples. Cambridge: Cambridge University Press. p 193-221.

Hoppa RD. 2000. Population Variation in Osteological Aging Criteria: An Example From the Pubic Symphysis. American Journal of Physical Anthropology 111:185-191.

Hoppa RD, and Vaupel JW. 2002. The Rostock Manifesto for Paleodemography: The Way from Stage to Age. In: Hoppa RD, and Vaupel JW, editors. Paleodemography Age Distributions from Skeletal Samples. Cambridge: Cambridge University Press. p 1-8.

IBM Corp. Released 2013 IBM SPSS Statistics for Windows, Version 22.0. Armonk, NY: IBM Corp

Igarashi Y, Uesu K, Wakebe T, and Kanazawa E. 2005. New Method for Estimation of Adult Skeletal Age at Death from the Morphology of the Auricular Surface of the Ilium. American Journal of Physical Anthropology 128:324-339.

İşcan MY, and Loth SR. 1986. Determination of Age from the Sternal Rib in White Males: A Test of the Phase Method. Journal of Forensic Sciences 31(1):122-132.

İşcan MY, and Loth SR. 1988. Assessment of Intercostal Variation on the Estimation of Age from the Sternal End of the Rib. 58th Annual Meeting of the American Association of Physical Anthropologists. San Diego, California.

İşcan MY, Loth SR, and Wright RK. 1984a. Age Estimation fron the Rib by Phase Analysis: White Males. Journal of Forensic Sciences 29(4):1094-1104.

İşcan MY, Loth SR, and Wright RK. 1984b. Metamorphosis at the Sternal Rib End: A New Method to Estimate Age at Death in White Males. American Journal of Physical Anthropology 65:147-156.

İşcan MY, Loth SR, and Wright RK. 1985. Age Estimation from the Rib by Phase Analysis: White Females. Journal of Forensic Sciences 30:853-863.

İşcan MY, Loth SR, and Wright RK. 1987. Racial Variation in the Sternal Extremity of the Rib and its Effect on Age Determination. Journal of Forensic Sciences 32(2):452-466.

Jackes M. 2000. Building the Bases for Paleodemographic Analysis: Adult Age Estimation. In: Katzenberg MA, and Saunders SR, editors. Biological Anthropology of the Human Skeleton. New York: Wiley-Liss. p 417-466.

Jackes MK. 1985. Pubic Symphysis Age Distribution. American Journal of Physical Anthropology 68:281-299.

Katz D, and Suchey JM. 1986. Age Determination of the Male Os Pubis. American Journal of Physical Anthropology 69:427-435.

Kemkes-Grottenthaler A. 1996. Sterbealterbestimmung anhand des ektokranialen Nahtverschlusses: Eine Evaluierung der Meindl-Lovejoy-Methode. Rechtsmedizin 6:177-184.

Kemkes-Grottenthaler A. 2002. Aging through the Ages: Historical Perspectives on Age Indicator Methods. In: Hoppa RD, and Vaupel JW, editors. Paleodemography Age Distributions from Skeletal Samples. Cambridge: Cambridge University Press. p 48-72.

Kerley ER. 1965. The Microscopic Determination of Age in Human Bone. American Journal of Physical Anthropology 23:149-164.

Kerley ER. 1969. Age Determination of bone Fragments. Journal of Forensic Sciences 14:59-67.

Key CA, Aiello LC, and Molleson T. 1994. Cranial Suture Closure and its Implications for Age Estimation. International Journal of Osteoarchaeology 4:193-207.

Kimmerle EH, Konigsberg LW, Jantz RL, and Baraybar JP. 2008a. Analysis of Age-at-Death Estimation Through the Use of Pubic Symphyseal Data. Journal of Forensic Sciences 53(3):558-568.

Kimmerle EH, Prince DA, and Berg GE. 2008b. Inter-Observer Variation in Methodologies Involving the Pubic Symphysis, Sternal Ribs, and Teeth. Journal of Forensic Sciences 53(3):594-600.

Klippel WE, and Synstelien JA. 2007. Rodents as Taphonomic Agents: Bone Gnawing by Brown Rats and Gray Squirrels. Journal of Forensic Sciences 52(4):765-773.

Konigsberg LW, and Frankenberg SR. 1992. Estimation of Age Structure in Anthropological Demography. American Journal of Physical Anthropology 89:235-256.

Konigsberg LW, Herrmann NP, Wescott DJ, and Kimmerle EH. 2008. Estimation and Evidence in Forensic Anthropology: Age-at-Death. Journal of Forensic Sciences 53(3):541-557.

Kunos CA, Simpson SW, Russell KF, and Hershkovitz I. 1999. First Rib Metamorphosis: Its Possible Utility for Human Age-at-Death Estimation. American Journal of Physical Anthropology 110:303-323.

Lamendin H, Baccino E, Humbert JF, Tavernier JC, Nossintchouk RM, and Zerilli A. 1992. A Simple Technique for Age Estimation in Adult Corpses: the Two Criteria Dental Method. Journal of Forensic Sciences 37:1373-1379.

Lottering N, MacGregor DM, Meredith M, Alston CL, and Gregory LS. 2013. Evaluation of the Suchey–Brooks Method of Age Estimation in an Australian Subpopulation using Computed Tomography of the Pubic Symphyseal Surface. American Journal of Physical Anthropology 150:386-399.

Lovejoy CO, Meindl RS, Mensforth RP, and Barton TJ. 1985a. Multifactorial Age Determination of Skeletal Age at Death: A Method and Blind Tests of its Accuracy. American Journal of Physical Anthropology 68:1-14.

Lovejoy CO, Meindl RS, Pryzbeck TR, and Mensforth RP. 1985b. Chronological Metamorphosis of the Auricular Surface of the Ilium: A New Method of Determining Adult Age at Death. American Journal of Physical Anthropology 68:15-28.

Mann RW, Jantz RL, Bass WM, and Willey PS. 1991. Maxillary Suture Obliteration: A Visual Method for Estimating Skeletal Age. Journal of Forensic Sciences 36:781-791.

Mann RW, Symes SA, and Bass WM. 1987. Maxillary Suture Obliteration: Aging the Human Skeleton Based on Intact or Fragmentary Maxilla. Journal of Forensic Sciences 32(1):148-157.

Masset C. 1989. Age Estimation on the Basis of Cranial Sutures. In: Iscan MY, editor. Age Markers in the Human Skeleton. Springfield: Charles C. Thomas. p 71-103.

McKern TW, and Stewart TD. 1957. Skeletal Age Changes in Young American Males. Natick: Quartermaster Research and Development Command.

Meindl RS, and Lovejoy CO. 1985. Ectocranial Suture Closure: A Revised Method for the Determination of Skeletal Age at Death Based on the Lateral-Anterior Sutures. American Journal of Physical Anthropology 68:57-66.

Meindl RS, Lovejoy CO, and Mensforth RP. 1983. Skeletal Age at Death: Accuracy of Determination and Implications for Paleodemography. Human Biology 55:73-87.

Meindl RS, Lovejoy CO, Mensforth RP, and Walker RA. 1985. A Revised Method of Age Determination Using the Os Pubis, with a Review and Tests of Accuracy of Other Current Methods of Pubic Symphyseal Aging. American Journal of Physical Anthropology 68:29-45.

Millán MS, Rissech C, and Turbón D. 2013. A test of Suchey-Brooks (pubic symphysis) and Buckberry-Chamberlain (auricular surface) method on an identified Spanish sample: paleodemographic implications. Journal of Archaeological Science 40:1743-1751.

Moore-Jansen PH, and Jantz RL. 1986. A Computerized Skeletal Data Bank for Forensic Anthropology. Knoxville, TN: Department of Anthropology, University of Tennessee.

Müller H-G, Love B, and Hoppa RD. 2002. Semiparametric Method for Estimating Paleodemographic Profiles from Age Indicator Data. American Journal of Physical Anthropology 117(1):1-14.

Murray KA, and Murray T. 1991. A test of the auricular surface aging technique. Journal of Forensic Sciences 36:1162-1169.

Nagaoka T, and Hirata K. 2008. Demographic structure of skeletal populations in historic Japan: A new estimation of adult age-at-death distributions based on the auricular surface of the ilium. Journal of Archaeological Science 35:1370-1377.

Osborne DN, Simmons TL, and Nawrocki SP. 2004. Reconsidering the auricular surface as an indicator of age at death. Journal of Forensic Sciences 49:1-7.

Prince DA, and Konigsberg LW. 2008. New Formulae for Estimating Age-at-Death in the Balkans Utilizing Lamendin's Dental Technique and Bayesian Analysis. Journal of Forensic Sciences 53(3):578-587.

Prince DA, and Ubelaker DH. 2002. Application of Lamendin's adult dental aging technique to a diverse skeletal sample. Journal of Forensic Sciences 47:107-116.

Ritz S, Schütz HW, and Peper C. 1993. Postmortem estimation of age at death based on aspartic acid racemization in dentin: its applicability for root dentin. International Journal of Legal Medicine 105:289-293.

Saunders SR, Fitzgerald C, Rogers T, Dudar C, and McKillop H. 1992. A Test of Several Methods of Skeletal Age Estimation Using a Documented Archaeological Sample. Canadian Society of Forensic Science Journal 25:97-117.

Scheuer L, and Black S. 2000. Developmental Juvenile Osteology. New York: Elsevier Academic Press.

Schmitt A. 2004. Age-at-Death Assessment using the Os Pubis and the Auricular Surface of the Ilium: a Test on an Identified Asian Sample. International Journal of Osteoarchaeology 14:1-6.

Schmitt A, Murail P, Cunha E, and Rouge D. 2002. Variability of the pattern of aging on the human skeleton: Evidence from bone indicators and implications on age at death estimation. Journal of Forensic Sciences 47:1203-1209.

Singer R. 1953. Estimation of age from cranial suture closure. Journal of Forensic Medicine 1:52-59.

Snow CC. 1983. Equations for Estimating Age at Death from the Pubic Symphysis: A Modification of the McKern-Stewart Method. Journal of Forensic Sciences 28(4):864-870.

Suchey JM. 1979. Problems in the Aging of Females Using the Os pubis. American Journal of Physical Anthropology 51:467-470.

Todd TW. 1920. Age changes in the Pubic Bone. I. The Male White Pubis. American Journal of Physical Anthropology 3:285-339.

Todd TW. 1921. Age Changes in the Pubic Bone. American Journal of Physical Anthropology 4(1):1-77.

Todd TW, and Lyon DW. 1924. Endocranial Suture Closure: Its Progress and Age Relationship Part I. Adult Males of White Stock. American Journal of Physical Anthropology 7(3):325-384.

Todd TW, and Lyon DW. 1925a. Cranial Suture Closure: Its Progress and Age Relationship Part II. Ectocranial Closure in Adult Males of White Stock. American Journal of Physical Anthropology 8(1):23-45.

Todd TW, and Lyon DW. 1925b. Cranial Suture Closure: Its Progress and Age Relationship Part III. Endocranial Closure in Adult Males of Negro Stock. American Journal of Physical Anthropology 8(1):47-71.

Todd TW, and Lyon DW. 1925c. Cranial Suture Closure: Its Progress and Age Relationship Part IV. Ectocranial Closure in Adult Males of Negro Stock. American Journal of Physical Anthropology 8(2):149-168.

Welcker H. 1862. Untersuchungen über Wachsthum und Bau des Menschlichen Schädels. Leipzig. xvi-148 p.

Welcker H. 1866. Kraniologische Mittheilungen. Arch f Anthrop 1:89-160.

Appendices

Appendix A:
Variable Definitions

Table A.1: Pubic symphysis variables with references and coding.

Reference	Age Indicator	# of Categories	Variable Code
Todd 1920	Pubic Symphysis Morphology	10	M1 Left
Todd 1921	Pubic Symphysis Morphology	10	M2 Left
Brooks and Suchey 1990	Pubic Symphysis Morphology	6	M3 Left
Hanihara and Suzuki 1978	Horizontal Ridges	4	M4 X1 L
	Pubic Tubercle	2	M4 X2 L
	Lower End	3	M4 X3 L
	Dorsal Margin	4	M4 X4 L
	Superior Ossific Nodule	3	M4 X5 L
	Ventral Bevelling	4	M4 X6 L
	Symphyseal Rim	2	M4 X7 L
McKern and Stewart 1957	Dorsal Demiface	6	M5 C1 L
	Ventral Rampart	6	M5 C2 L
	Symphyseal Rim	6	M5 C3 L
Gilbert and McKern 1973	Dorsal Demiface	6	M6 C1 L
	Ventral Rampart	6	M6 C2 L
	Symphyseal Rim	6	M6 C3 L
Chen et al. 2008	Ridges and Furrows	5	M7 A L
	Ridges of pubic tubercle	3	M7 B L
	Lower Extremity	4	M7 C L
	Ventral Rampart	4	M7 D L
	Ossific Nodule	3	M7 E L
	Dorsal Margin	4	M7 F L
	Ventral Bevelling	3	M7 G L
	General Macroscopic Changes	3	M7 H L
	Bone Density	4	M7 I L
Boldsen et al. 2002	Symphyseal Relief	6	M8 C1 L
	Symphyseal Texture	4	M8 C2 L
	Superior Apex	4	M8 C3 L
	Ventral Symphyseal Margin	7	M8 C4 L
	Dorsal Symphyseal Margin	5	M8 C5 L

Table A.2: Auricular surface variables with references and coding.

Reference	Age Indicator	# of Stages	Variable Code
Lovejoy et al. 1985b	Auricular Surface Morphology	8	M11 Left
Boldsen et al. 2002	Superior Demiface Topography	3	M8 C6 L
	Inferior Demiface Topography	3	M8 C7 L
	Superior Surface Morphology	5	M8 C8 L
	Apical Surface Morphology	5	M8 C9 L
	Inferior Surface Morphology	5	M8 C10 L
	Inferior Surface Texture	3	M8 C11 L
	Superior Posterior Iliac Exostoses	6	M8 C12 L
	Inferior Posterior Iliac Exostoses	6	M8 C13 L
	Posterior Iliac Exostoses	3	M8 C14 L
Buckberry and Chamberlain 2002	Transverse Organization	5	M9 C1 L
	Surface Texture	5	M9 C2 L
	Microporosity	3	M9 C3 L
	Macroporosity	3	M9 C4 L
	Apical Changes	3	M9 C5 L
Igarashi et al. 2005	Wide Groove	2	M10 WG L
	Striation	2	M10 ST L
	Roughness	2	M10 RO L
	Flatness	2	M10 FL L
	Smoothness	2	M10 SM L
	Fine Granularity	2	M10 FG L
	Coarse Granularity	2	M10 CG L
	Sparse Porosity	2	M10 SP L
	Dense Porosity	2	M10 DP L
	Dull Rim	2	M10 DR L
	Lipping	2	M10 LP L
	Tuberosity	2	M10 TB L
	Bony Bridge	2	M10 BB L

Table A.3: Rib morphology variables with references and coding.

Reference	Age Indicator	# of Categories	Variable Code
Isçan et al. 1984a	Sternal Rib End Morphology	9	M12 R4 L
Isçan et al. 1985	Sternal Rib End Morphology	9	M13 R4 L
Isçan et al. 1984b	Pit Depth	6	M14 R4 C1 L
	Pit Shape	6	M14 R4 C2 L
	Rim and Wall Configuration	6	M14 R4 C3 L

Table A.4: Cranial suture variables with references and coding.

Reference	Age Indicator [(B)ilateral]	# of Categories	Variable Code
Meindl and Lovejoy 1985	Midlambdoid (B)	4	M15 S1 L
	Lambda	4	M15 S2
	Obelion	4	M15 S3
	Anterior Sagittal	4	M15 S4
	Bregma	4	M15 S5
	Midcoronal (B)	4	M15 S6 L
	Pterion (B)	4	M15 S7 L
	Sphenofrontal (B)	4	M15 S8 L
	Inferior Sphenotemporal (B)	4	M15 S9 L
	Superior Sphenotemporal (B)	4	M15 S10 L
	Incisive	4	M15 S11
	Anterior Median Palatine	4	M15 S12
	Posterior Median Palatine	4	M15 S13
	Transverse Palatine	4	M15 S14
Boldsen et al. 2002	Coronal Pterica	5	M8 S1 L
	Sagittal Obelica	5	M8 S2
	Lambdoidal asterica (B)	5	M8 S3 L
	Zygomaticomaxillary (B)	5	M8 S4 L
	Interpalatine (Posterior)	5	M8 S5

Table A.5: Redefined variables with their origins

Variable Code	Original	# of Stages	Source
M4 X32 L	M4 X3 L	4	Hanihara And Suzuki 1978
M4 X62 L	M4 X6 L	4	
M7 Gx L	M7 G L	3	Chen et al. 2007
M7 Ga L	M7 G L	continuous	
M7 G2 L	M7 G L	3	
M8 C15 L	M8 C14 L	3	Boldsen et al. 2002
M10 TB2 L	M10 TB L	2	Igarashi et al. 2005

Table A.6: Variables created out of combining two single sex methods

Male Method	Female Method	Feature	Variable Code	# of Stages
Todd 1920	Todd 1921	Pubic Symphysis Morphology	M1.5 Left	10
McKern and Stewart 1957	Gilbert and McKern 1973	Dorsal Demiface	M5.5 C1 L	6
		Ventral Rampart	M5.5 C2 L	6
		Symphyseal Rim	M5.5 C3 L	6
Işçan et al. 1984a	Işçan et al. 1985	Sternal Rib End Morphology	M12.5 R4 L	9

Appendix B:
Age Method Definitions

This appendix contains the definitions of age indicator stages as they were taken from the literature. They are presented in order of variable coding as well as age indicators starting with pubic symphysis methods, followed by auricular surface methods, rib methods, and cranial suture methods. The author defined variables are given after the literature definitions of the method from which the new variable was derived. The three methods that have new variable definitions added on to their descriptions are Hanihara and Suzuki (1978), Chen et al. (2008), and Igarashi et al. (2005).

Pubic Symphysis

The following descriptions are taken directly from Todd (1920: 301-314).

Phase 1: Symphysial surface rugged; traversed by horizontal ridges separated by well marked grooves; no ossific (epiphyseal) nodules fusing with the surface; no definite delimiting margin; no definition of extremities.

Typical adolescent ridge and furrow formation with no sign of margins and no ventral beveling.

Phase 2: Symphysial surface still rugged, traversed by horizontal ridges, the grooves between which are, however, becoming filled near the dorsal limit with a new formation of finely textured bone. This formation begins to obscure the hinder extremities of the horizontal ridges. Ossific (epiphysial) nodules fusing with the upper symphysial face may occur; dorsal limiting margin begins to develop; no delimitation of extremities; foreshadowing of ventral bevel.

Foreshadowing of ventral beveling with slight indication of dorsal margin.

Phase 3: Symphyseal face shows progressive obliteration of ridge and furrow system; commencing formation of the dorsal plateau; presence of fusing ossific (epiphyseal) nodules; dorsal margin gradually becoming more defined; beveling as a result of ventral rarefaction becoming rapidly more pronounced; no delimitation of extremities.

Progressive obliteration of ridge and furrow system with increasing definition of dorsal margin and commencement of ventral rarefaction (beveling).

Phase 4: Great increase of ventral beveled area; corresponding diminution of ridge and furrow formation; complete definition of dorsal margin through the formation of the dorsal plateau; commencing delimitation of lower extremity.

Completion of definite dorsal margin, rapid increase of ventral rarefaction and commencing delimitation of lower extremity.

Phase 5: Little or no change in symphysial face and dorsal plateau except that sporadic and premature attempts at the formation of a ventral rampart occur; lower extremity, like the dorsal margin, is increasing in clearness of definition; commencing formation of upper extremity with or without the intervention of a bony (epiphysial) nodule.

Commencing formation of upper extremity with increasing definition of lower extremity and possibly sporadic attempts at formation of ventral rampart.

Phase 6: Increasing definition of extremities; development and practical completion of ventral rampart; retention of granular appearance of symphysial face and ventral aspect of pubis; absence of lipping of symphysial margin.

Development and practical completion of ventral rampart with increasing definition of extremities.

Phase 7: Changes in symphysial face and ventral aspect of pubis consequent upon diminishing activity; commencing bony outgrowth into attachments of tendons and ligaments, especially the gracilis tendon and sacro-tuberous ligament.

Changes in symphysial face and ventral aspect of pubis consequent upon diminishing activity, accompanied by bony outgrowths into pelvic attachments of tendons and ligaments.

Phase 8: Symphysial face generally smooth and inactive; ventral surface of pubis also inactive; oval outline complete or approximately complete; extremities clearly defined; no distinct "rim" to symphysial face; no marked lipping of either dorsal or ventral margin.

Smoothness and inactivity of symphysial face and ventral aspect of pubis. Oval outline and extremities clearly defined but no "rim" formation or lipping.

Phase 9: Symphysial face presents a more or less marked rim; dorsal margin uniformly lipped; ventral margin irregularly lipped.

Development of "rim" on symphysial face with lipping of dorsal and ventral margins.

Phase 10: Symphysial face eroded and showing erratic ossification; ventral border more or less broken down; disfigurement increases with age.

Erosion of and erratic, possibly pathological osteophytic growth on symphysial face with breaking down of ventral margin.

The following descriptions are taken directly from Todd (1921: 27-34).

Phase 1: Symphysial surface rugged, traversed by horizontal ridges separated by well marked grooves, there being no distinction in size between the upper and the lower ridges and the whole pattern being more delicate than in the male. No bony (epiphysial) nodules fusing with the surface. No definite delimiting margin. No definition of extremities.

Phase 2: Symphysial face still rugged. The horizontal grooves are becoming filled near their dorsal limit with new finely textured bone. Bone (epiphysial) nodules fusing with upper symphysial face. Dorsal delimiting margin begins to develop. No delimitation of extremities. Ventral bevel commencing.

Phase 3: Symphyseal face shows progressive obliteration of ridge and furrow system. Commencing formation of dorsal platform. Possible presence of bony nodules. Dorsal margin becoming more defined and sharply lipped. Ventral bevel more pronounced. Extremities not delimited.

Phase 4: Great increase of ventral beveled area. Corresponding diminution of ridge and furrow formation. Complete definition of dorsal margin through the formation of the dorsal platform. Commencing delimitation of lower extremity.

Phase 5: Relatively small change in symphysial face and dorsal platform except for sporadic efforts a t the formation of a ventral rampart. Dorsal margin increasingly clearly defined and more sharply lipped. Lower extremity better defined. Upper extremity forming with or without the intervention of a bony (epiphysial) nodule.

Phase 6: Increasing definition of extremities. Development and practical completion of ventral rampart. Retention to a small degree of granular appearance of symphysial face indicating that activity has not yet quite ceased. Failure of ventral aspect of pubis adjacent to ventral rampart to become transformed into a compact surface. Because of this the rampart is more or less undermined. Retention of pectinate outline of dorsal margin and in slight degree of ridge and furrow system. No lipping of ventral margin and no increased lipping of dorsal margin.

Phase 7: Slight changes in symphysial face and marked changes in ventral aspect consequent upon diminishing activity. No formation of symphysial rim. No ossification of tendinous and ligamentous attachments.

Phase 8: Symphysial face and ventral aspect of pubic bone generally smooth and inactive. Oval outline complete. Extremities clearly defined. No distinct "rim" to symphysial face. No lipping of ventral or increased lipping of dorsal margin. Development of ossification in tendinous and ligamentous attachments especially those of sacro-tuberous ligament and gracilis muscle.

Phase 9: Symphysial face presents a more or less marked rim. No lipping of ventral and no further lipping of dorsal margin. No secondary erosion or rarefaction.

Phase 10: Ventral margin eroded over a greater or less extent of its length, continuing somewhat into the symphysial face. No increased lipping. Disfigurement only occasional and slight.

The following descriptions are taken directly from Brooks and Suchey (1990: 232-233).

Phase 1: Symphyseal face has a billowing surface (ridges and furrows) which usually extends to include the pubic tubercle. The horizontal ridges are well-marked and ventral beveling may be commencing. Although ossific nodules may occur on the upper extremity, *a key to the recognition of this phase is the lack of delimitation of either extremity (upper or lower).*

Phase 2: The symphyseal face may still show ridge development. *The face has commencing delimitation of lower and/or upper extremities occurring with or without ossific nodules.* The ventral rampart may be in beginning phases as an extension of the bony activity at either or both extremities.

Phase 3: Symphyseal face shows lower extremity and *ventral rampart in process of completion*. There can be a continuation of fusing ossific nodules forming the upper extremity along the ventral border. Symphyseal face is smooth or can continue to show distinct ridges. Dorsal plateau is complete. Absence of lipping of symphyseal dorsal margin; no bony ligamentous outgrowths.

Phase 4: Symphyseal face is generally fine grained although remnants of the old ridge and furrow system may still remain. *Usually the oval outline is complete at this stage, but a hiatus can occur in upper ventral rim*. Pubic tubercle is fully separated from the symphyseal face by definition of upper extremity. The symphyseal face may have a distinct rim. Ventrally, bony ligamentous outgrowths may occur on inferior portion of pubic bone adjacent to symphyseal face. If any lipping occurs it will be slight and located on the dorsal border.

Phase 5: *Symphyseal face is completely rimmed with some slight depression of the face itself, relative to the rim*. Moderate lipping is usually found on the dorsal border with more prominent ligamentous outgrowths on the ventral border. There is little or no rim erosion. Breakdown may occur on superior ventral border.

Phase 6: *Symphyseal face may show ongoing depression as rim erodes*. Ventral ligamentous attachments are marked. In many individuals the pubic tubercle appears as a separate bony knob. The face may be pitted or porous, giving an appearance of disfigurement with the ongoing process of erratic ossification. Crenulations may occur. The shape of the face is often irregular at this stage.

The following information is taken directly from Hanihara and Suzuki (1978: 234).

Variable: Horizontal ridges and furrows (X_1)

Stage 1: Distinct

Stage 2: Furrows become shallow

Stage 3: Trace

Stage 4: No longer visible

Variable: Pubic tubercle (X_2)

Stage 1: Attached by cartilage

Stage 2: United

Variable: Lower end (X_3)

Stage 1: Indistinct

Stage 2: Narrow ridge

Stage 3: Broad ridge

Variable: Dorsal margin (X_4)

Stage 1: None

Stage 2: Interrupted narrow ridge

Stage 3: Narrow ridge over full length

Stage 4: Broad ridge

Variable: Superior ossific nodule (X_5)

Stage 1: None

Stage 2: Present

Stage 3: No longer visible

Variable: Ventral bevelling (X_6)

Stage 1: None

Stage 2: Incomplete

Stage 3: Completed over full length

Stage 4: Upper part no longer visible

Variable: Symphyseal rim (X_7)

Stage 1: Incomplete

Stage 2: Whole symphyseal surface bordered by a broad rim

Author's redefined variables derived from Hanihara and Suzuki (1978):

Variable: Lower extremity (X32):

Score 1: No edge, no rim, no protrusion

Score 2: No rim present at lower extremity but a noticeable edge or bump is present

Score 3: The rim presents a sharp edge like a cliff dropping off onto the ischiopubic ramus

Score 4: dull edge that is rounded or sloped or is indistinct

Variable: Ventral Bevelling (X62)

Score 1: None

Score 2: Incomplete

Score 3: Viewed ventrally there are furrows, rarefaction, or sulcus visible

Score 4: Viewed ventrally the bone is continuously flat, or shows uniform erosion or rarefaction

The following descriptions are taken directly from McKern and Stewart (1957: 75-79).

Component 1: Dorsal Plateau

Score 0: Dorsal margin absent.

Score 1: A slight margin formation first appears in the middle third of the dorsal border.

Score 2: The dorsal margin extends along entire dorsal border.

Score 3: Filling in of grooves and resorption of ridges to form a beginning plateau in the middle third of the dorsal demi-face.

Score 4: The plateau, still exhibiting vestiges of billowing, extends over most of the dorsal demi-face.

Score 5: Billowing disappears completely and the surface of the entire demi-face becomes flat and slightly granulated in texture.

Component 2: Ventral Rampart

Score 0: Ventral beveling is absent.

Score 1: Ventral beveling is present only at superior extremity of ventral border.

Score 2: Bevel extends inferiorly along ventral border.

Score 3: The ventral rampart begins by means of bony extensions from either of both of the extremities.

Score 4: The rampart is extensive but gaps are still evident along the earlier ventral border, most evident in the upper two-thirds.

Score 5: The rampart is complete.

Component 3: Symphyseal Rim

Score 0: The symphyseal rim is absent.

Score 1: A partial dorsal rim is present, usually at the superior end of the dorsal margin, it is round and smooth in texture and elevated above the symphyseal surface.

Score 2: The dorsal rim is complete and the ventral rim is beginning to form. There is no particular beginning site.

Score 3: The symphyseal rim is complete. The enclosed symphyseal surface is finely grained in texture and irregular or undulating in appearance.

Score 4: The rim begins to break down. The face becomes smooth and flat and the rim is no longer round but sharply defined. There is some evidence of lipping on the ventral edge.

Score 5: Further breakdown of the rim (especially along superior ventral edge) and rarefaction of the symphyseal face. There is also disintegration and erratic ossification along the ventral rim.

The following descriptions are taken directly from Gilbert and McKern (1973: 33-34).

Component 1: Dorsal demi-face

Score 0: Ridges and furrows very distinct, ridges are billowed, dorsal margin undefined.

Score 1: Ridges begin to flatten, furrows to fill in, and a flat dorsal margin begins in mid-third of demi-face.

Score 2: Dorsal demi-face spreads ventrally, becomes wider as flattening continues dorsal margin extends superiorly and inferiorly.

Score 3: Dorsal demi-face is quite smooth, margin may be narrow or indistinct from face.

Score 4: Demi-face becomes complete and unbroken, is broad and very fine grained, may exhibit vestigial billowing.

Score 5: Demi-face becomes pitted and irregular through rarefaction.

Component 2: Ventral Rampart

Score 0: Ridges and furrows very distinct. The entire demi-face is beveled up toward the dorsal demi-face.

Score 1: Beginning inferiorly, the furrows of the ventral demi-face begin to fill in, forming an expanding beveled rampart, the lateral edge of which is a distinct, curved line extending the length of the symphysis.

Score 2: Fill in of furrows and expansion of demi-face continue from both superior and inferior ends, rampart spreads laterally along its ventral edge.

Score 3: All but about one-third of the ventral demi-face is filled in with fine grained bone.

Score 4: The ventral rampart presents a broad, complete, fine grained surface from the pubic crest to the inferior ramus.

Score 5: Ventral rampart may begin to break down, assuming a very pitted and perhaps cancellous appearance through rarefaction.

Component 3: Symphyseal Rim

Score 0: The rim is absent.

Score 1: The rim begins in the mid-third of the dorsal surface.

Score 2: The dorsal part of the symphyseal rim is complete.

Score 3: The rim extends from the superior and inferior ends of the symphysis until all but about one-third of the ventral aspect is complete.

Score 4: The symphyseal rim is complete.

Score 5: Ventral margin of dorsal demi-face may break down so that gaps appear in the rim, or it may round off so that there is no longer a clear dividing line between the dorsal demi-face and the ventral rampart.

The following descriptions are taken directly from Chen and colleagues (2008: 37-40).

Component A: Ridges and furrows on the symphyseal surface

Score 0: ridges and furrows alternate distinctly

Score 1: the furrows fill in and ridges and furrows alternate indistinctly

Score 2: the bone substance has a granular look with low, blunt ridges and shallow furrows

Score 3: the surface becomes flat and fine-textured, and/or again become more granular

Score 4: ridges and furrows disappear entirely and the surface becomes pitted and eroded

Component B: Ridges of the pubic tubercle

Score 0: ridges on pubic tubercle completed

Score 1: ridges on pubic tubercle almost gone

Score 2: ridges on pubic tubercle completely disappeared

Component C: Lower extremity of the symphyseal surface

Score 0: no appearance of lower extremity

Score 1: appearance of dividing line between symphyseal surface and inferior ramus of pubis

Score 2: forming a “V” angle

Score 3: atrophy or disappearance of "V" angle

Component D: Ventral rampart

Score 0: no appearance of ventral rampart

Score 1: local ventral rampart

Score 2: fully developed ventral rampart

Score 3: ventral rampart becomes wider or nodular in superior portion

Component E: Ossific nodule

Score 0: no appearance of ossific nodule

Score 1: appearance of ossific nodule

Score 2: fusion and disappearance of ossific nodule

Component F: Dorsal margin

Score 0: no appearance of dorsal margin

Score 1: edged margin without a plateau

Score 2: forming a plateau and lip-like thickness and extension in superior part of dorsal margin

Score 3: middle destruction or generalized atrophy of dorsal margin

Component G: Ventral beveling

Score 0: no appearance of ventral beveling

Score 1: clear-edged margin with a right angle between ventral beveling and the symphyseal surface

Score 2: ventral beveling becomes flat in the lower portion or disappears

Component H: General macroscopic changes of the symphyseal surface

Score 0: prominence of symphyseal surface

Score 1: irregular surface

Score 2: flatness, or fovea inferior with clear periphery

Component I: Bone density of the symphyseal surface

Score 0: ridged or rough with no ridge

Score 1: smooth, dense and solid

Score 2: concavo-convex or with dense pores on surface

Score 3: big pits and/or loss of density

Author's redefined variables derived from Chen et al. (2008):

Component Ga: Ventral Bevel Angle:

Continuous Variable measuring the angle between the line lying flat on the pubic symphyseal surface and the line along the ventral surface of the pubic bone. The vertex of the angle is commonly on the ventral portion of the symphyseal rim, or in younger individuals where this rim would form later. The measurement is taken at the superior-inferior midpoint of the symphysis.

Component Gx: Ventral Bevel Angle

Score 0: The ventral surface exhibits furrows, the ventral rampart is partially complete but not finished.

Score 1: the ventral rampart is complete and the ventral bevel angle measures between 90^0 and 60^0.

Score 2: the ventral rampart is complete and the ventral bevel angle measures between 0^0 and 59^0.

Component G2: Ventral Bevel

Score 0: the surface shows furrows without any evidence of beveling

Score 1: The bevel is visible to any extent including through rarefaction after ventral rampart has formed.

Score 2: The bevel is obscured by the rampart or continuous dense bone on the ventral surface. The ventral surface looks uniform, either as smooth dense bone, or exhibiting rarefaction over the whole ventral aspect.

The following descriptions are taken directly from Boldsen and colleagues (2002: 97-100).

Component 1: Symphyseal relief

Score 1: *Sharp billows*: At least half of the entire symphyseal face is covered with sharply crested billows. These billows consist of distinct ridges separated by deep furrows, and they extend completely across the symphyseal face. The low parts of furrows cut deeply into the ventral and dorsal margins of the symphyseal face. In some specimens, great vertical relief is accompanied by rounded, not sharp, crests on billows. Symphyseal faces are scored as having *Sharp Billowing* if the distance between the high and low points of adjacent ridges and furrows is 3 mm or more. This stage has only been noted in bones from teenagers, especially young ones.

Score 2: *Soft, deep billows*: At least half of the symphyseal face, typically the dorsal demiface, is covered with softly crested to flat billows separated by deep furrows. There is no obvious filling of furrows with bone.

Score 3: *Soft, shallow billows*: Much of the symphyseal surface, typically the dorsal demiface, is covered by shallow, but clearly visible and discrete, billows. Remnants of the ridge-and-furrow system are clearly visible. The billows extend most, or sometimes all, of the way across the face.

Score 4: *Residual billows*: Billows blend into one another, and they form an important element of the surface, but they are much less pronounced than in the previous categories. The subtle billows do not fulfill the criteria for the previous categories. Two or more billows that conform to the residual category must be present. They typically extend only part way across the symphyseal face, usually no more than one-half the width of the face.

Score 5: *Flat*: Over one-half of the symphyseal face is flat or slightly recessed. It sometimes presents a pebbly appearance because of the presence of numerous small, flat, pillows of bone. The rest of the symphysis does not indicate a billow score (i.e., no more than one discrete billow is present).

Score 6: *Irregular*: Over one-half of the symphyseal surface is markedly irregular because of pitting, which is sometimes deep, often accompanied by small, sharp exostoses scattered thickly across the face. Occasionally the entire surface of what is otherwise a flat face is covered by the small knobs of bone (here pitting is largely absent). None of the criteria used to define earlier symphyseal stages is fulfilled.

Component 2: Symphyseal texture

Score 1: *Smooth*: Most, or all, of the dorsal demiface is covered by fine-grained or smooth bone.

Score 2: *Coarse*: More than one-third of the dorsal demiface consists of coarse-textured bone.

Score 3: *Microporosity*: More than one-third of the dorsal demiface is covered by bone that has a porous appearance. The overall impression is of numerous, closely packed, pin-pricks.

Score 4: *Macroporosity*: More than one-third of the dorsal demiface is marred by generally closely spaced, deep pits, which are 0.5 mm or more in diameter. They collectively give the face an irregular, porous appearance.

Component 3: Superior apex

Score 1: *No protuberance*: The surface of the cranial end of the symphyseal face displays deep to shallow billowing. There is no evidence of a raised bony protuberance.

Score 2: *Early protuberance*: A distinct knob of bone is present in the cranial end of the symphyseal face. This rounded bony protuberance is clearly differentiated from the immediately adjacent symphyseal face (i.e., the symphyseal face and the ventral beveled area, which is often present).

Score 3: *Late protuberance*: The cranial end of the symphyseal face immediately anterior to the midline is raised somewhat above the rest of the articulation surface. The margins of the protuberance are poorly defined, creating a raised area that is more completely integrated with the remainder of the symphyseal face than in the previous category. The raised area should not be confused with a narrow elevated rim defining the cranial end of the symphyseal face. In some specimens the cranial part of the face may be more of less isolated by breakdown pitting but these faces should not be coded as belonging to the *Late protuberance* stage. That is to say, a raised area of bone must be visible on a rather smooth symphyseal face.

Score 4: *Integrated*: There is no raised area of bone on the cranial end of the symphyseal face. The symphyseal face is flat and it has a smooth or pitted appearance. The area where the protuberance was located is fully integrated with the rest of the symphyseal face.

Component 4: Ventral symphyseal margin

Score 1: *Serrated*: The ventral edge of the pubic symphysis is irregular because of an uninterrupted extension of the ridges and furrows typical of pronounced billowing.

Score 2: *Beveled*: There is a distinct flattening (or loss) of billows in the ventral portion of the pubic symphysis. The beveling generally begins in the superior part of the ventral demiface. It must extend over one-third or more of the ventral margin to be scored as present.

Score 3: *Rampart formation*: The ventral rampart refers to a distinct outgrowth of bone defining the ventral margin of the symphyseal face. In this stage the bony rampart is incomplete, it does not extend along the entire ventral edge, and usually some ridges and furrows on the symphyseal face can be followed uninterrupted to the ventral edge of the symphysis. Often the remnants of billows can be seen dipping below the partially formed ventral rampart, which looks like a roll of gum laid across a shallowly furrowed surface. An incomplete ventral rampart often extends inferiorly from the bony protuberance defining the cranial end of the face. An incomplete ventral rampart can also extend superiorly from the caudal end of the pubic symphysis. In many specimens there is a gap in the middle one-third of the ventral margin where bony ramparts from the ends of the symphysis have not yet met. Specimens in an early formation stage can have one or more bony knobs, which are often located in the middle one-third of the ventral margin. These knobs occur with or without the bony extensions from the cranial and caudal ends of the symphysis. A well-developed bony protuberance at the cranial end of the face that lacks a distinct inferiorly oriented extension of bone (the rampart) should not be coded as *Rampart formation*; i.e., the cranially located knob is not alone sufficient to score the ventral rampart as being present.

Score 4: *Rampart completion I*: The ventral rampart is complete. There is, however, a shallow sulcus that extends for much of the length of the ventral surface of the pubis immediately lateral to the ventral edge of the symphyseal face. This groove is a residual feature related to rampart extension over the original symphyseal surface. Occasionally there is gap in the ventral rampart, usually in the superior half of the ventral margin; the ventral rampart, however, is otherwise completely formed. The flat pubic symphyseal surface, which extends uninterrupted from its dorsal to ventral margins, contrasts with the typically furrowed *Rampart formation* stage.

Score 5: *Rampart completion II*: The ventral rampart is complete. There is no shallow sulcus as described in *Rampart completion I*. Occasionally there is a gap in the ventral rampart, usually in the superior half of the ventral margin; however, the ventral rampart is otherwise completely formed. The flat pubic symphyseal surface, which extends uninterrupted from its dorsal to ventral margins, contrasts with the furrowed appearance of the typical *Rampart formation* stage.

Score 6: *Rim*: There is a narrow bony rum on the ventral rampart that demarcates a generally flat or irregular symphyseal face. The ventral rim can be incomplete or complete, but it must be at least 1 cm long and readily identifiable as a distinct raised ridge bordering a slightly recessed symphyseal face. The presence of a rim meeting the length criterion is sufficient to score the pubis a being in the *Ventral rim* stage, regardless of the ventral rampart configuration.

Score 7: *Breakdown*: The ventral aspect of the symphyseal face shows signs of breakdown, which takes the form of pitting and an erosion of part of the ventral margin. To be scored as present, the breakdown of the ventral margin must exceed 1 cm (either in one place or when two or more areas of erosion are combined).

Component 5: Dorsal symphyseal margin

Score 1: *Serrated*: The dorsal edge of the pubic symphysis is irregular because of an uninterrupted extension of the ridges and furrows typical of pronounced billowing.

Score 2: *Flattening incomplete*: There is a well-defined flattened area at least 1 cm long, usually in the superior part of the dorsal demiface, where the articular surface meets the dorsal surface of the pubis. Some residual billowing is present that produces an undulating dorsal edge, which is not as extreme as that found in the *Serrated* category. This undulating edge is typically found along the inferior dorsal margin.

Score 3: *Flattening complete*: There is a complete, or virtually complete, well-defined area of flattening where the symphyseal face meets the dorsal surface of the pubis. Occasionally there will be a small area at the inferior end of the dorsal margin that still retains an undulating appearance.

Score 4: *Rim*: There is a narrow bony rim at least 1 cm long demarcating a generally flat or irregular face. The dorsal rim can be incomplete or complete, but it must be readily identifiable as a raised ridge bordering a slightly recessed symphyseal face. It generally appears first along the superior part of the dorsal margin.

Score 5: *Breakdown*: The dorsal aspect of the symphyseal face shows signs of breakdown, which takes the form of pitting and erosion of the dorsal margin. To be scored as present, the breakdown of the dorsal margin must exceed 1 cm in length (either in one place or when two or more areas of erosion are combined). Antemortem destruction of the dorsal margin attributable to large parity pits that undercut the symphyseal face can occur in females, but it is not considered breakdown, and it often results in this feature being unscorable.

Auricular Surface

The following descriptions are taken directly from Lovejoy and colleagues (1985: 21-27).

Phase 1: Billowing and very fine granularity

The surface displays fine granular texture and marked transverse organization. There is no retro-auricular activity, apical activity, or porosity. The surface appears youthful because of broad and well-defined billows, which impart the definitive transverse organization. Billows are well-defined and cover most of the surface. Any subchondral defects are smooth-edged and rounded.

Phase 2: Reduction of billowing, but retention of youthful appearance

Changes from the previous phase are not marked and are mostly reflected in slight to moderate loss of billowing, with replacement by striae. There is no apical activity, porosity, or retroauricular activity. The surface still appears youthful owing to marked transverse organization. Granulation is slightly more coarse.

Phase 3: General loss of billowing, replacement by striae, and distinct coarsening of granularity

Both faces are largely quiescent with some loss of transverse organization. Billowing is much reduced and replaced by (definite) striae. The surface is more coarsely and recognizably granular than in previous phase, with no significant changes at apex. Small areas of microporosity may appear. Slight retroauricular activity may occasionally be present. In general, coarse granulation supercedes and replaces billowing.

Phase 4: Uniform coarse granularity

Both faces are coarsely and uniformly granulated, with marked reduction of both billowing and striae, but striae may still be present under close examination. Transverse organization is present but poorly defined. There is some activity in the retroauricular area but this is usually slight. Minimal changes are seen at the apex, microporosity is slight, and there is no macroporosity. This is the primary period of uniform granularity.

Phase 5: Transition from coarse granularity to dense surface; this may take part over islands of the surface of one or both faces.

No billowing is seen. Striae may be present but very vague. The face is still partially (coarsely) granular and there is a marked loss of transverse organization. Partial densification (which may occur in islands) of the surface with commensurate loss of grain is present along with slight to moderate activity in the retroauricular area. Occasional macroporosity is seen, but this is not typical. Slight changes are usually present at apex. Some increase in microporosity is seen, depending upon the degree of densification. The primary feature is the transition from a granular to a dense surface.

Phase 6: Completion of densification with complete loss of granularity

Significant loss of granulation is seen in most specimens, with replacement by dense bone. No billows or striae are present. Changes at apex are slight to moderate but are almost always present. There is a distinct tendency for the surface to become dense. No transverse organization is evident. Most or all of any microporosity is lost to densification process. There is increased irregularity of margins with moderate retroauricular activity and little or no macroporosity.

Phase 7: Dense irregular surface of rugged topography and moderate to marked activity in periauricular areas

This is a further elaboration of previous stage, in which marked surface irregularity becomes paramount feature. Topography, however, shows no transverse or other form of organization. Moderate granulation is occasionally retained, but is usually lost during previous phase and is generally absent. No striae or billows are present. The inferior face generally is lipped at inferior terminus, so as to extend beyond the body of the innominate bone. Apical changes are almost invariable and may be marked. Increasing irregularity of margins is seen. Macroporosity is present in some cases but it is not requisite. Retroauricular activity is moderate to marked in most cases.

Phase 8: Breakdown with marginal lipping, macroporosity, increased irregularity, and marked activity in periauricular areas.

The paramount feature is a nongranular, irregular surface, with distinct signs of subchondral destruction. No transverse organization is seen and there is a definitive absence of any youthful criteria. Macroporosity is present in about one-third of all cases. Apical activity is usually marked but is not requisite for this age category. Margins become dramatically irregular and lipped, with typical degenerative joint change. The retroauricular area becomes well defined with profuse osteophytes of low to moderate relief.

The following descriptions are taken directly from Boldsen and colleagues (2002: 101-103).

Component 6/7: Superior/inferior demiface topography

Score 1: *Undulating*: The surface is slightly undulating, especially in a superior to inferior direction. The rise and fall of the bone surface is often best detected by feel. There is no centrally located area of elevated bone (the *Median elevation*). Surface billows superimposed on the wavy joint surface give it a somewhat hummocky appearance. The superior demiface, especially the most cranial part of it, is typically flatter than the inferior demiface.

Score 2: *Median elevation*: In the middle of the demiface there is a broad raised area where the middle part of the joint is elevated above the rest of the surface. This bony elevation is flanked anteriorly and posteriorly by one or two long low areas. The

elevated area takes the form of an elongated ridge, particularly in the inferior demiface, with the long axis paralleling the main orientation of the joint. This ridge need not occupy the entire length of the joint surface.

Score 3: *Flat to irregular*: The surface is essentially flat or recessed, a result of marginal lipping, or it is irregular, a result of a degeneration of the joint or the formation of low pillow-like exostoses. Sometimes the inferior demiface has a slight curve to it so the inferior portion is located somewhat laterally to the superior part, a result of the joint conforming to the general shape of the ilium in this area. In such instances, the articulation surface does not have the softly rounded, wavy appearance of the *Undulating* category.

Component 8/9/10: Superior/apical/inferior surface morphology

Score 1: *Billows over >2/3 of the surface*: Low rounded and typically narrow ridges separated by furrows, which have rounded bases, are clearly identifiable. The ridge surfaces are curved from the depths of the furrows completely across their crests. Most or all of the billowing is oriented roughly anterior to posterior, and individual furrows sometimes run across much, or all, of the face. The billowing covers most (>2/3) of the auricular surface (i.e., it is a dominant element of the surface).

Score 2: *Billows over 1/3 to 2/3 of the surface*: About one-half of the surface is covered by billows.

Score 3: *Billows over <1/3 of the surface*: Billows are a noticeable, but minor, component of the joint surface. The rest of the surface is flat or bumpy.

Score 4: *Flat*: The auricular surface is flat.

Score 5: *Bumps*: Most, or all, of the auricular surface is covered by low, rounded areas of raised bone, much like little irregular pillows. Part of the surface may be flat, but over half of it is bumpy.

Component 11: Inferior surface texture

This part of the joint surface is 1 cm long, as measured in a superior to inferior direction. Its lowermost point is a line defined by the margin of the greater sciatic notch on either side of the sacroiliac joint.

Score 1: *Smooth*: Most or all of the bone comprising the auricular surface exhibits a smooth to slightly granular appearance.

Score 2: *Microporosity*: At least one-half of the surface has a porous appearance with the apertures being less than 0.5 mm in diameter. The surface appears to be covered with numerous closely spaced pinpricks.

Score 3: *Macroporosity*: At least one half of the surface is porous, and most or all of the apertures exceed 0.5 mm in diameter.

Component 12/13: Superior/inferior posterior iliac exostoses

The two areas examined are located on the medial surface of the posterior ilium where ligaments attach during life. The superior area is superior to the sacroiliac joint surface; .e., to a line that passes from the anterior superior iliac spine to the most superior point of the

joint surface (the superior angle), and on through the posterior part of the ilium. The inferior area is located inferior to that line. It is immediately posterior to the middle of the sacroiliac joint; i.e., it lies behind the most anteriorly projecting part of the joint's posterior margin. Exostoses appear on all but the bones of the youngest adults (with rare exceptions), and they tend to be clustered together to form nicely defined and easily identifiable patches of rough bone.

Score 1: *Smooth*: The iliac surface is flat to slightly raised, but the surface is smooth. That is to say, it shows no evidence of round to sharp bony elevations. At most there are a few isolated and very small exostoses.

Score 2: *Rounded exostoses*: Definite but low exostoses with rounded crests dominate the scoring areas.

Score 3: *Pointed exostoses*: Sharply pointed but still low exostoses dominate the scoring areas.

Score 4: *Jagged exostoses*: The scoring areas have a jagged appearance because of the presence of high round to sharp exostoses.

Score 5: *Touching exostoses*: There is a pronounced outgrowth of bone with a relatively flat top, which is usually roughly oval, where the raised part of the ilium meets the sacrum.

Score 6: *Fused*: The ilium and sacrum are fused together.

Component 14: Posterior iliac exostoses

The area that is examined is the medial side of the ilium bordered posteriorly by the iliac crest, anteriorly by the sacroiliac auricular surface, superiorly by a slightly raised area often surmounted by exostoses (superior posterior iliac exostoses), and inferiorly by a similarly raised area also typically covered by exostoses (inferior posterior iliac exostoses). As opposed to the areas where the superior and posterior iliac exostoses are located, the part of the ilium of interest here is much less likely to have enough bony projections to be counted as present (i.e., rounded or pointed).

Score 1: *Smooth*: The area posterior to the sacroiliac joint is smooth, except for the two areas coded as the superior and inferior posterior iliac exostoses. Surfaces interrupted by isolated projections of bone, either rounded or sharp, are still considered as smooth. Such exostoses typically occur on all but the youngest adults, yet much of the original smooth iliac surface is retained.

Score 2: *Rounded exostoses*: Low, rounded, bony projections cover the entire bone surface posterior to the sacroiliac joint, except for a ca. 1 cm band of smooth bone immediately adjacent to the posterior edge of the joint. The entire surface is rough because little, if any, of the original smooth iliac surface remains. The exostoses are normally lower than the superior and inferior posterior iliac exostoses.

Score 3: *Pointed spicules*: Low, pointed bony projections cover the entire bone surface posterior to the sacroiliac joint, except for a ca. 1 cm band of smooth bone immediately adjacent to the posterior edge of the joint. The entire surface is rough

because little, if any, of the original smooth iliac surface remains. The exostoses are normally lower than the superior and inferior posterior iliac exostoses.

Author's Redefined Variable derived from Boldsen et al. (2002):

Component 15: The area scored for this variable is the kidney-shaped area right behind the sacroiliac joint, between the sacroiliac joint and the retroauricular surface. It is scored with the same definitions as component 14 (above):

Score 1: *Smooth*: The area posterior to the sacroiliac joint is smooth, except for the two areas coded as the superior and inferior posterior iliac exostoses. Surfaces interrupted by isolated projections of bone, either rounded or sharp, are still considered as smooth. Such exostoses typically occur on all but the youngest adults, yet much of the original smooth iliac surface is retained.

Score 2: *Rounded exostoses*: Low, rounded, bony projections cover the entire bone surface posterior to the sacroiliac joint, except for a ca. 1 cm band of smooth bone immediately adjacent to the posterior edge of the joint. The entire surface is rough because little, if any, of the original smooth iliac surface remains. The exostoses are normally lower than the superior and inferior posterior iliac exostoses.

Score 3: *Pointed spicules*: Low, pointed bony projections cover the entire bone surface posterior to the sacroiliac joint, except for a ca. 1 cm band of smooth bone immediately adjacent to the posterior edge of the joint. The entire surface is rough because little, if any, of the original smooth iliac surface remains. The exostoses are normally lower than the superior and inferior posterior iliac exostoses.

The following descriptions are taken directly from Buckberry and Chamberlain (2002; 233-234).

Component 1: Transverse organization

Score 1: 90% or more of surface is transversely organized

Score 2: 50-89% of surface is transversely organized

Score 3: 25-49% of surface is transversely organized

Score 4: Transverse organization is present on less than 25% of surface

Score 5: No transverse organization is present

Component 2: Surface texture

Score 1: 90% or more of surface is *finely granular*

Score 2: 50-89% of surface is *finely granular*; replacement of finely granular bone by coarsely granular bone in some areas; no dense bone is present

Score 3: 50% or more of surface is *coarsely granular*, but no dense bone is present

Score 4: *Dense bone* is present, but occupies less than 50% of surface; this may be just one small nodule of dense bone in very early stages

Score 5: 50% or more of surface is occupied by *dense bone*

Component 3: Microporosity

Score 1: No microporosity is present

Score 2: Microporosity is present on one demiface only

Score 3: Microporosity is present on both demifaces

Component 4: Macroporosity

Score 1: No macroporosity is present

Score 2: Macroporosity is present on one demiface only

Score 3: Macroporosity is present on both demifaces

Component 5: Apical changes

Score 1: Apex is sharp and distinct; auricular surface may be slightly raised relative to adjacent bone surface

Score 2: Some lipping is present at apex, but shape of articular margin is still distinct and smooth (shape of outline of surface at apex is a continuous arc)

Score 3: Irregularity occurs in contours of articular surface; shape of apex is no longer a smooth arc

The following descriptions are taken directly from Igarashi and colleagues (2005: 327).

Relief:

Wide groove (WG)	-Surface with wide transverse grooves having wide flat bottoms
Striation (ST)	-Surface with narrow transverse fine grooves that do not have flat bottoms and have V-shaped sections
Roughness (RO)	-Uneven surface without regular structures such as striations or wide grooves
Flatness (FL)	-Smooth surface without protrusions or depressions

Texture:

Smoothness (SM)	-Smooth even surface
Fine granularity (FG)	-Surface with depressions so shallow that no clear shadow is seen on surface under light from any direction
Coarse Granularity (CG)	-Surface with depressions deep enough so that clear shadows on surface can be seen. A surface with typical coarse granularity resembles surface of crepe de chine.
Sparse porosity (SP)	-Surface with pores that reach down to spongy bone and have a total area smaller than that of remaining surface of compact bone
Dense porosity (DP)	-Surface with many pores, total area of which is larger than that of remaining surface of compact bone

Hypertrophyed bony structures:

Dull rim (DR)	-Hypertrophied margin developed into a broad rim
Lipping (LP)	-Hypertrophied margin similar to a lip

Tuberosity (TB) -Nodes or spines often found just outside upper rear and lower front margins of auricular surfaces.

Bony bridge (BB) -Osteophyte in para-auricular region connecting sacrum and ilium, often found on upper part of sacroiliac joint.

Author's redefined variable (M10 TB2) derived from Igarashi et al. (2005):

Tuberosity 1 (TB) -Nodes or spines found at the lower front margin of auricular surface

Tuberosity 2 (TB2) -Nodes or spines found at the upper rear margin of the auricular surface, where the auricular surface ends and the retroauricular surface begins.

Rib Morphology

The following descriptions are taken directly from Isçan and colleagues (1984: 1096-1099).

Phase 0: The articular surface is flat or billowy with a regular rim and rounded edges. The bone itself is smooth, firm, and very solid.

Phase 1: There is a beginning amorphous indentation in the articular surface, but billowing may also still be present. The rim is rounded and regular. In some cases scallops may start to appear at the edges. The bone is still firm, smooth and solid.

Phase 2: The pit is now deeper and has assumed a V-shaped appearance formed by the anterior and posterior walls. The walls are thick and smooth with a scalloped or slightly wavy rim with rounded edges. The bone is firm and solid.

Phase 3: The deepening pit has taken on a narrow to moderately U-shape. Walls are still fairly thick with rounded edges. Some scalloping may still be present but the rim is becoming more irregular. The bone is still quite firm and solid.

Phase 4: Pit depth is increasing, but the shape is still a narrow to moderately wide U. The walls are thinner, but the edges remain rounded. The rim is more irregular with no uniform scalloping pattern remaining. There is some decrease in the weight and firmness of the bone, however, the overall quality of the bone is still good.

Phase 5: There is little change in pit depth, but the shape in this phase is predominantly a moderately wide U. Walls show further thinning and the edges are becoming sharp. Irregularity is increasing in the rim. Scalloping pattern is completely gone and has been replaced with irregular bony projections. The condition of the bone is fairly good, however, there are some signs of deterioration with evidence of porosity and loss of density.

Phase 6: The pit is noticeably deep with a wide U-shape. The walls are thin with sharp edges. The rim is irregular and exhibits some rather long bony projections that are frequently more pronounced at the superior and interior borders. The bone is noticeably lighter in weight, thinner, and more porous, especially inside the pit.

Phase 7: The pit is deep with a wide to very wide U-shape. The walls are thin and fragile with sharp, irregular edges and bony projections. The bone is light in weight and brittle with significant deterioration in quality and obvious porosity.

Phase 8: In this final phase the pit is very deep and widely U-shaped. In some cases the floor of the pit is absent or filled with bony projections. The walls are extremely thin, fragile, and brittle with sharp, highly irregular edges and bony projections. The bone is very lightweight, thin, brittle, friable, and porous. "Window" formation is sometimes seen in the walls.

The following descriptions are taken directly from Isçan and colleagues (1985: 855-858).

Phase 0: The articular surface is nearly flat with ridges or billowing. The outer surface of the sternal extremity of the rib is bordered by what appears to be an overlay of bone. The rim is regular with rounded edges, and the bone itself is firm, smooth, and very solid.

Phase 1: A beginning, amorphous indentation can be seen in the articular surface. Ridges or billowing may still be present. The rim is rounded and regular with a little waviness in some cases. The bone remains solid, firm, and smooth.

Phase 2: The pit is considerably deeper and has assumed a V-shape between the thick, smooth anterior and posterior walls. Some ridges or billowing may still remain inside the pit. The rim is wavy with some scallops beginning to form at the rounded edge. The bone itself is firm and solid.

Phase 3: There is only slight if any increase in pit depth, but the V-shape is wider, sometimes approaching a narrow U as the walls become a bit thinner. The still rounded edges now show a pronounced, regular scalloping pattern. At this stage, the anterior or posterior walls or both may first start to exhibit a central, semicircular arc of bone. The rib is firm and solid.

Phase 4: There is a noticeable increase in the depth of the pit, which now has a wide V- or narrow U-shape with, at times, flared edges. The walls are thinner but the rim remains rounded. Some scalloping is still present, along with the central arc; however, the scallops are not as well defined and the edges look somewhat worn down. The quality of the bone is fairly good but there is some decrease in density and firmness.

Phase 5: The depth of the pit stays about the same, but the thinning walls are flaring into a wider V- or U-shape. In most cases, a smooth, hard, plaque-like deposit lines at least part of the pit. No regular scalloping pattern remains and the edge is beginning to sharpen. The rim is becoming more irregular, but the central arc is still the most prominent projection. The bone is noticeably lighter in weight, density and firmness. The texture is somewhat brittle.

Phase 6: An increase in pit depth is again noted, and its V- or U-shape has widened again because of pronounced flaring at the end. The plaque-like deposit may still appear but is rougher and more porous. The walls are quite thin with sharp edges and an irregular rim. The central arc is less obvious and, in many cases, sharp points project from the rim of the sternal extremity. The bone itself is fairly thin and brittle with some signs of deterioration.

Phase 7: In this phase, the depth of the predominantly flared U-shaped pit not only shows no increase, but actually decreases slightly. Irregular bony growths are often seen extruding from the interior of the pit. The central arc is still present in most cases but is now accompanied by pointed projections, often at the superior and inferior borders, yet may be evidenced anywhere around the rim. The very thin walls have

irregular rims with sharp edges. The bone is very light, thin, brittle, and fragile, with deterioration most noticeable inside the pit.

Phase 8: The floor of the U-shaped pit in this final phase is relatively shallow, badly deteriorated, or completely eroded. Sometimes it is filled with bony growths. The central arc is barely recognizable. The extremely thin, fragile walls have highly irregular rims with very sharp edges, and often fairly long projections of bone at the inferior and superior borders. "Window" formation sometimes occurs in the walls. The bone itself is in poor condition--extremely thin, light in weight, brittle, and fragile.

The following descriptions are taken directly from Isçan and colleagues (1984: 148-149).

Component 1: Pit depth

Score 0: Flat to slightly billowy extremity with no indentation (pit) greater than 1.1 mm

Score 1: Definite pit formation with a depth ranging from 1.1 to 2.5 mm

Score 2: Pit depth ranging from 2.6 to 4.5 mm

Score 3: Pit depth ranging from 4.6 to 7.0 mm

Score 4: Pit depth ranging from 7.1 to 10.0 mm

Score 5: Pit depth of 10.1 mm or more

Component 2: Pit shape

Score 0: This stage is used for juvenile and adolescent specimens with no pit formation at the flat or billowy articular surface.

Score 1: A shallow, amorphous indentation (pit) is now present.

Score 2: Formation of a V-shaped pit with thick walls.

Score 3: The pit assumes a narrow U-shape with fairly thick walls.

Score 4: Wide U-shaped pit with thin walls.

Score 5: The pit is still a wide U-shape, yet deeper, more brittle, and poorer in texture with some disintegration of bone.

Component 3: Rim and wall configurations

Score 0: The 0 designation is for those specimens with a smooth regular rim and no wall formation.

Score 1: Beginning walls with a thick, smooth regular rim.

Score 2: Definitely visible walls that are thick and smooth with a scalloped or slightly wavy rim.

Score 3: A transitional stage between the regularity in stage 2 and the irregularity in stage 4. The scalloped edges are disappearing and the walls are thinning, yet the walls remain fairly sturdy without significant deterioration in the texture of the bone.

Score 4: The rim is becoming sharper and increasingly irregular with more frequent bony projections often most pronounced at the cranial and caudal margins of the rib. the

walls show further thinning and are less sturdy with noticeable deterioration in texture.

Score 5: The texture shows extreme friability and porosity. The rim is very sharp, brittle and highly irregular with long bony projections. Occasionally, as the depth of the pit increases, windows are formed in areas where the walls are not complete.

Cranial Sutures

The following descriptions are taken directly from Meindl and Lovejoy (1985; 58-60).
Score a small (1 cm) length of a suture or a specific site. Ignore other activity close to the site.

Score 0: Open

There is no evidence of any ectocranial closure at the site

Score 1: Minimal Closure

Some closure has occurred. This score is given for any minimal to moderate closure, i.e., from a single bony bridge across the suture to about 50% synostosis at the site.

Score 2: Significant Closure

There is a marked degree of closure but some portion of the site is still not completely fused.

Score 3: Complete Obliteration

The site is completely fused.

The sites that were scored are midlambdoid, lambda, obelion, anterior sagittal, bregma, midcoronal, pterion, sphenofrontal, inferior sphenotemporal, superior sphenotemporal, incisive, anterior median palatine, posterior median palatine, transverse palatine.

The following descriptions are taken directly from Boldsen and colleagues (2002; 103-104).
Ectocranial sutures to be scored are the Coronal pterica, Sagittal obelica, Lambdoidal asterica, Zygomaticomaxillary, and Interpalatine (Posterior portion of median palatine)

Score 1: Open

The suture is visible along its entire length, and there is a noticeable gap between the bones.

Score 2: Juxtaposed (not scorable for interpalatine suture)

The suture is visible along its entire length, but the suture is narrow because the bones are tightly juxtaposed. If bony bridges are present they are rare and very small (<1mm), sometimes with a trace of the original suture still evident.

Score 3: Partially obliterated

The suture is partially obscured. There is no trace of the original suture in the bony bridges.

Score 4: Punctuated

Only remnants of the sutures are present. These remnants appear as scattered small points or grooves, each no more than 2 mm long.

Score 5: Obliterated

There is no evidence of a suture.

Appendix C:
Supplemental statistical output

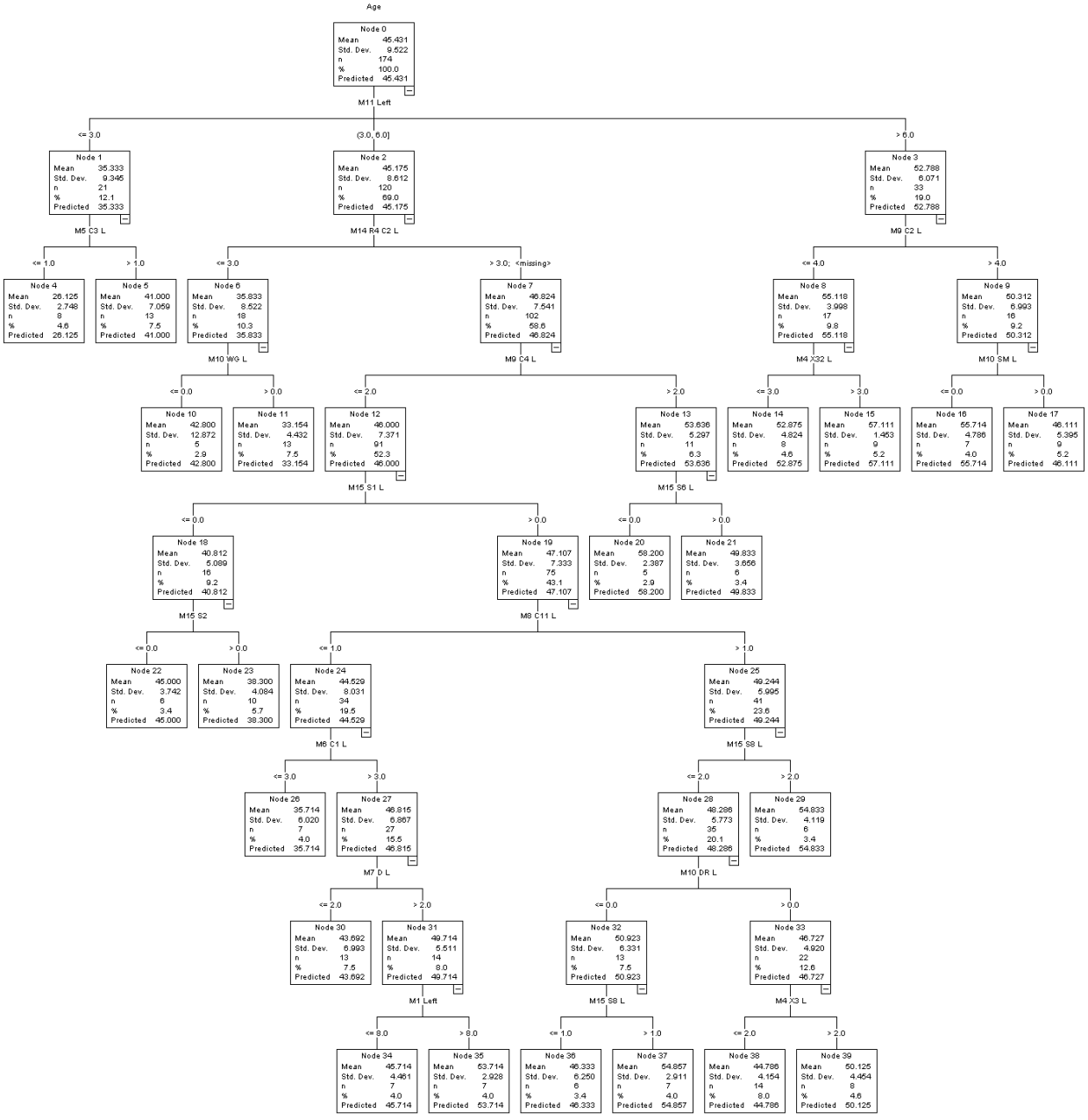


Figure C.1: CHAID 10/5 with all variables excluding observations 81 and 116

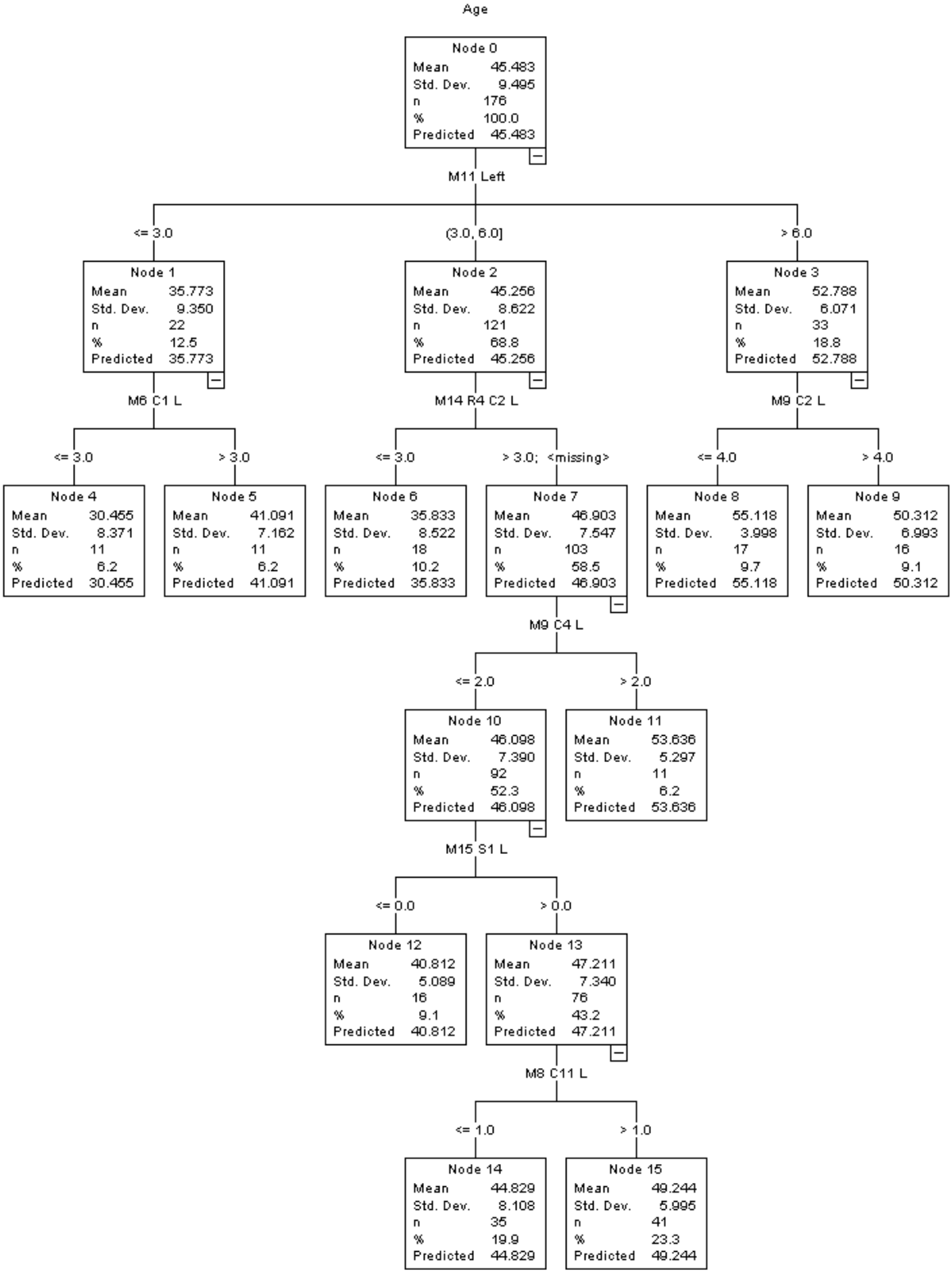


Figure C.2: CHAID 20/10 with all variables and observations

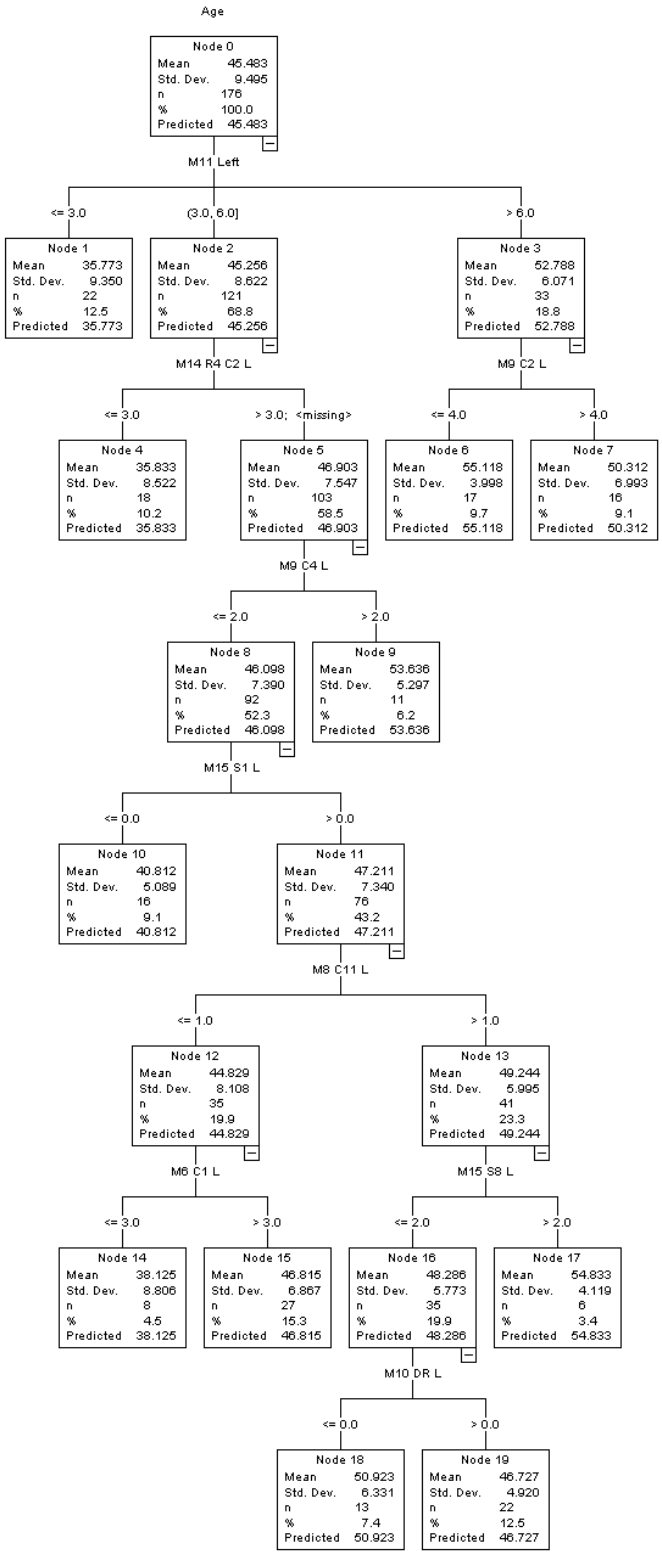


Figure C.3: CHAID 30/5 with all variables and observations

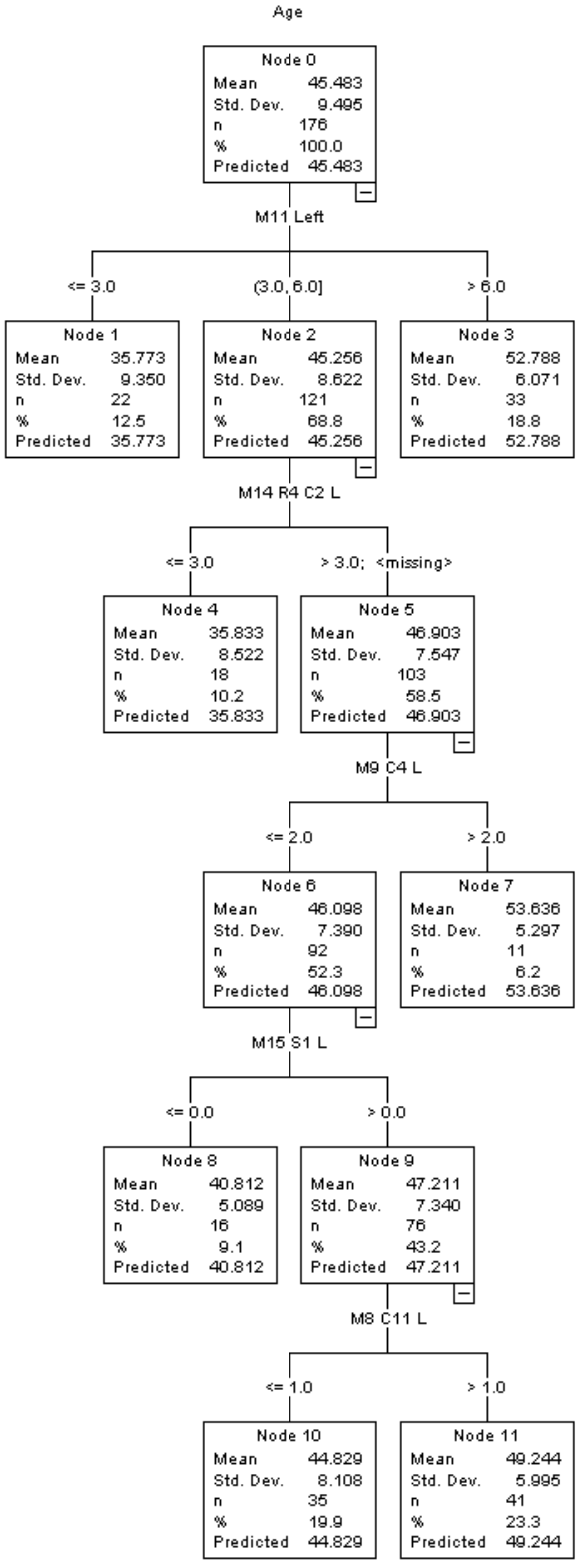


Figure C.4: CHAID 40/10 with all variables and observations

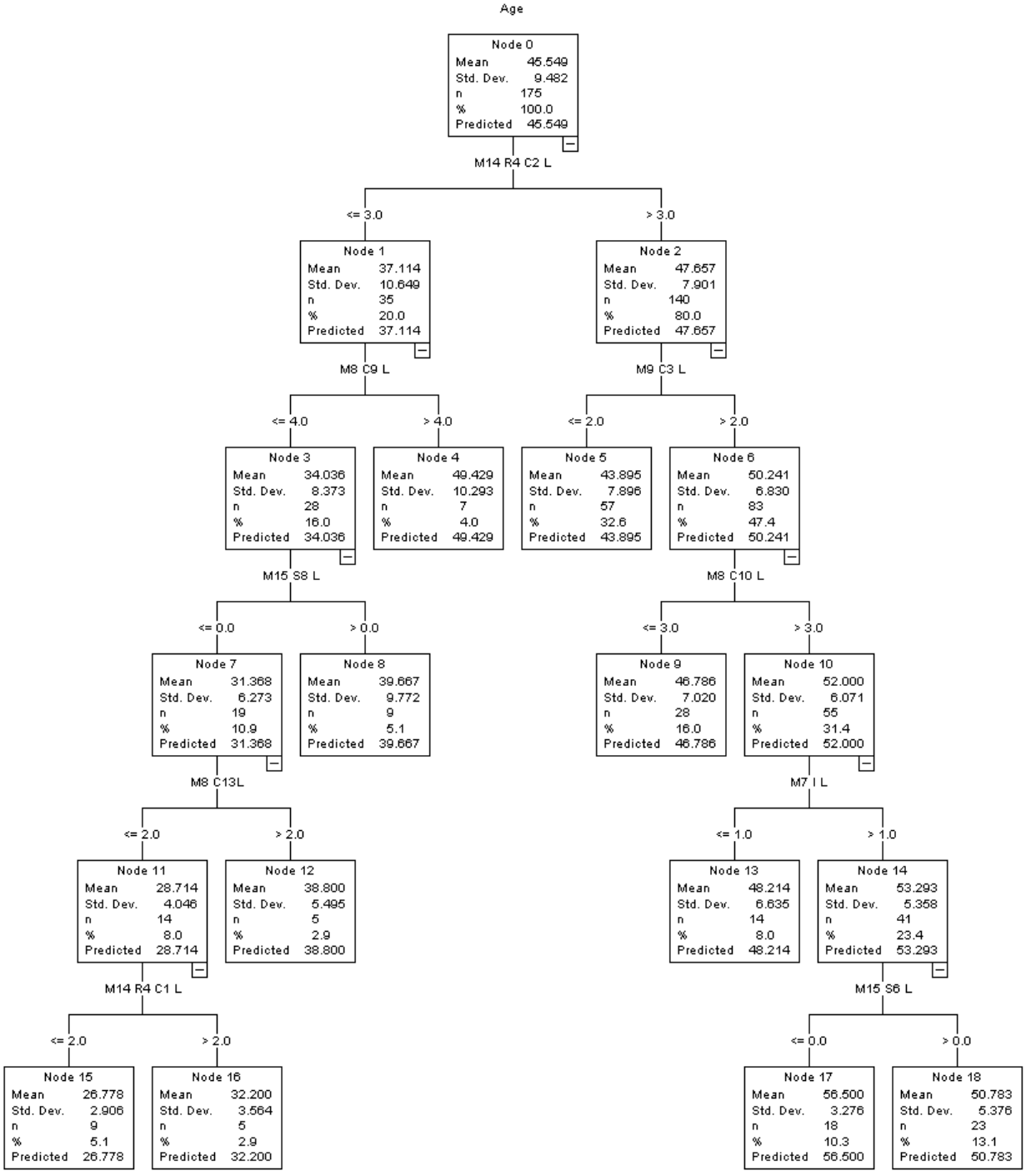


Figure C.5: CRT 10/5 with all variables and without observation 15

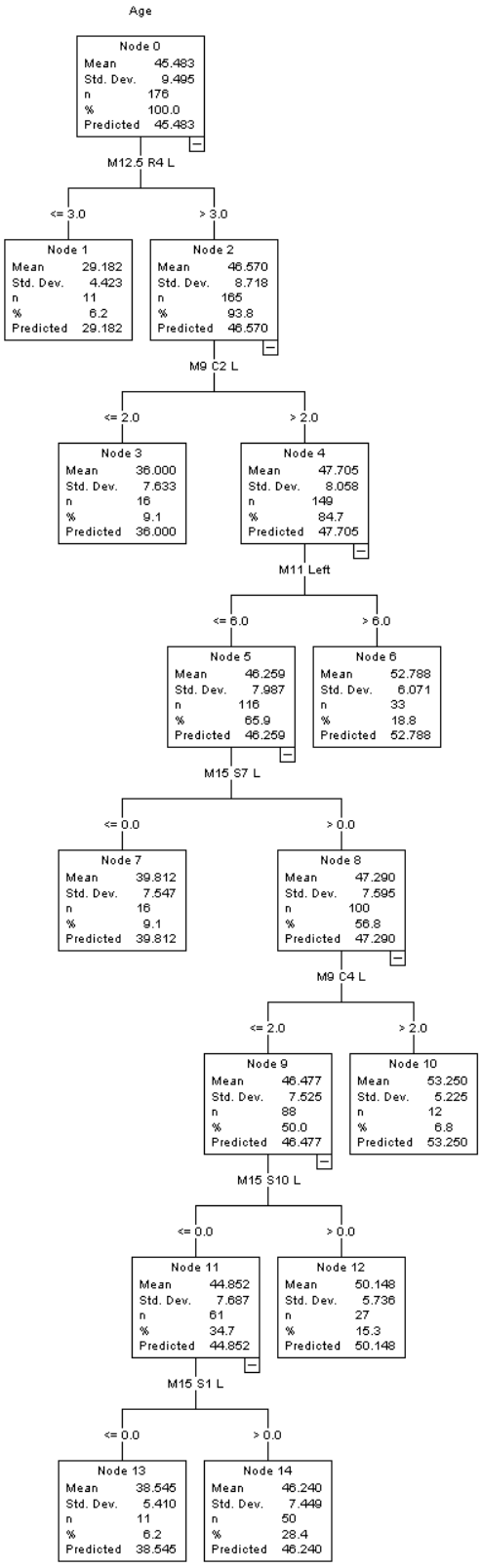


Figure C.6: CRT 50/10 with all variables and observations

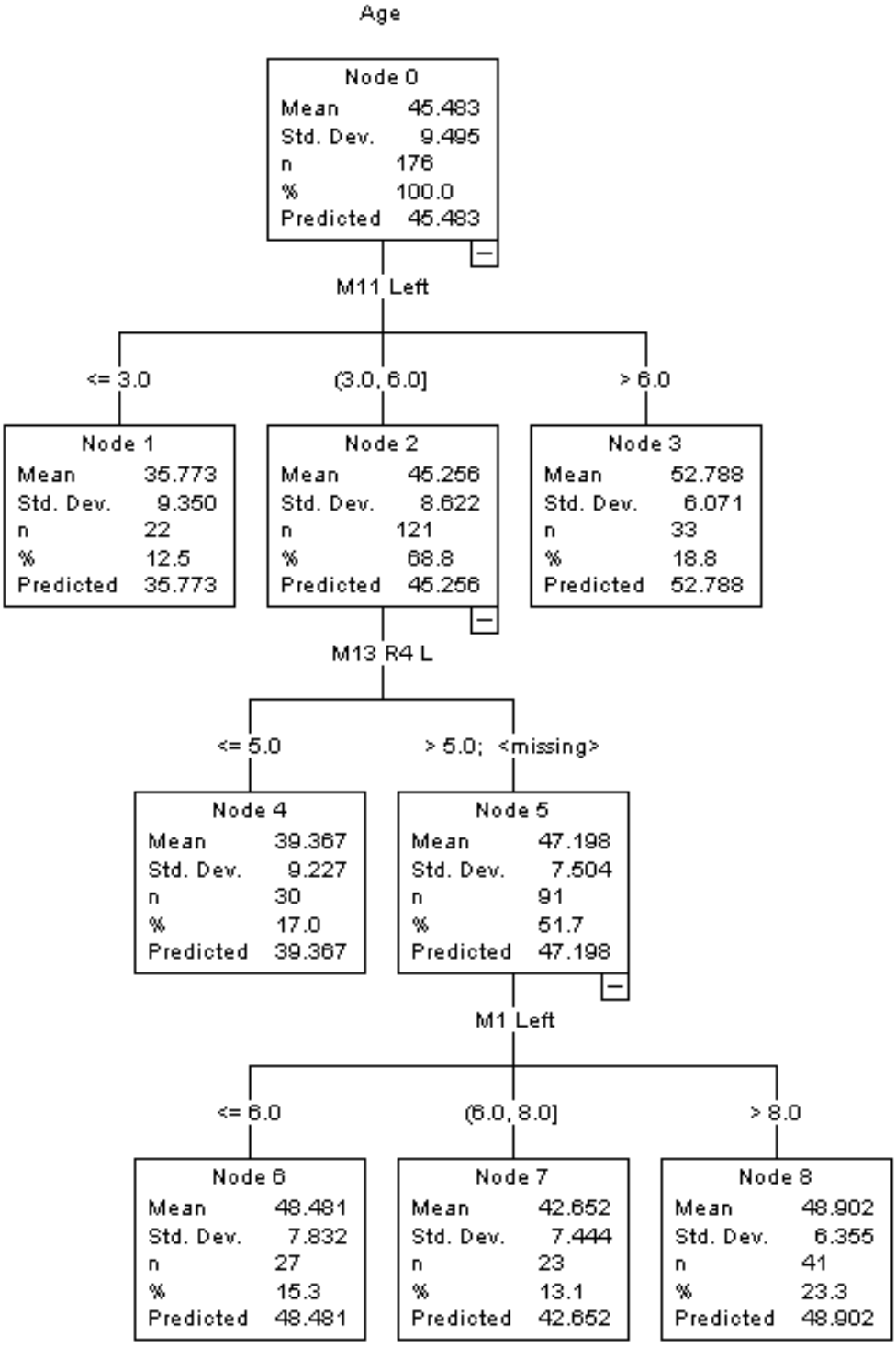


Figure C.7: CHAID 30/10 with all observations and only author's standard variables

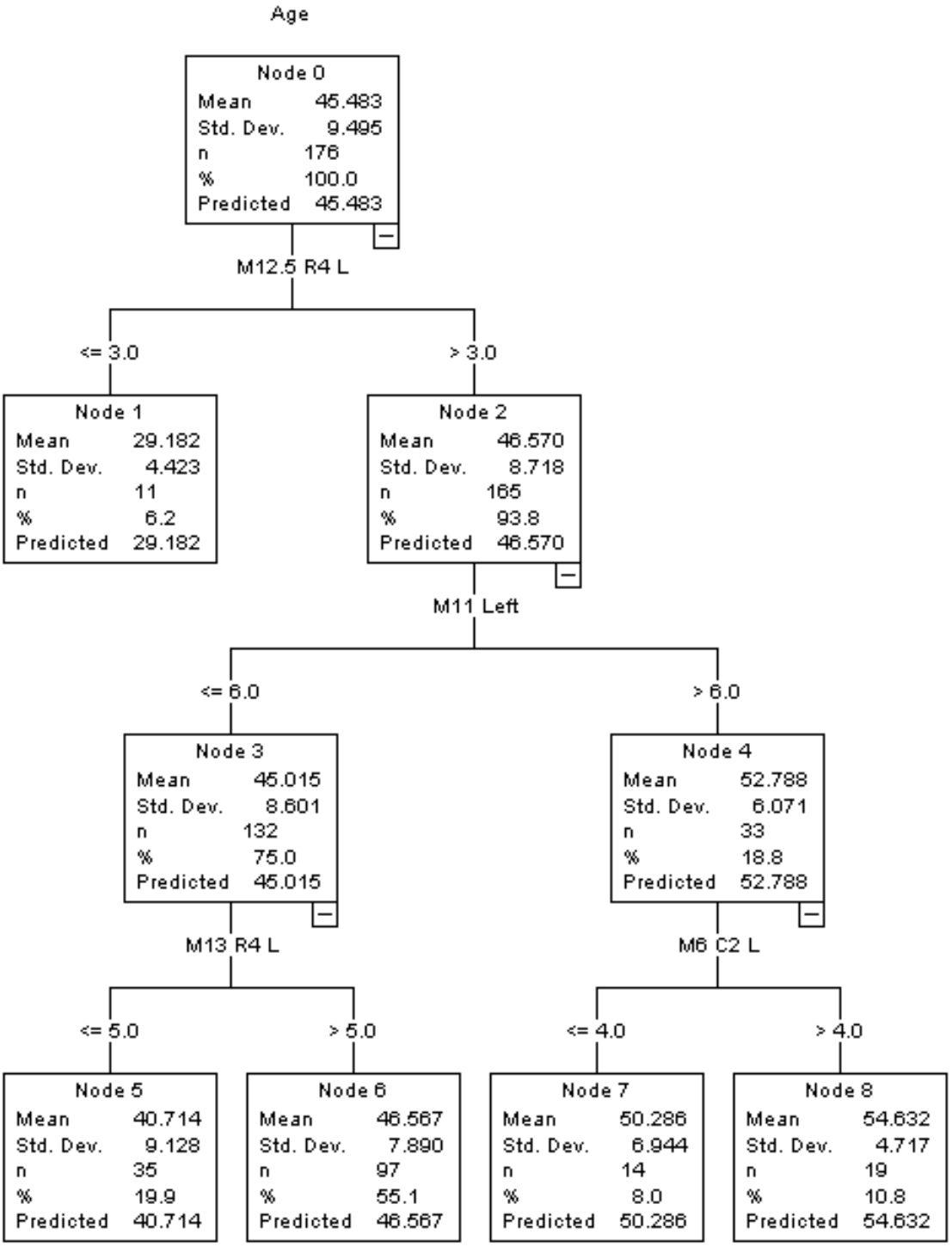


Figure C.8: CRT 30/10 with all observations and only author's standard variables

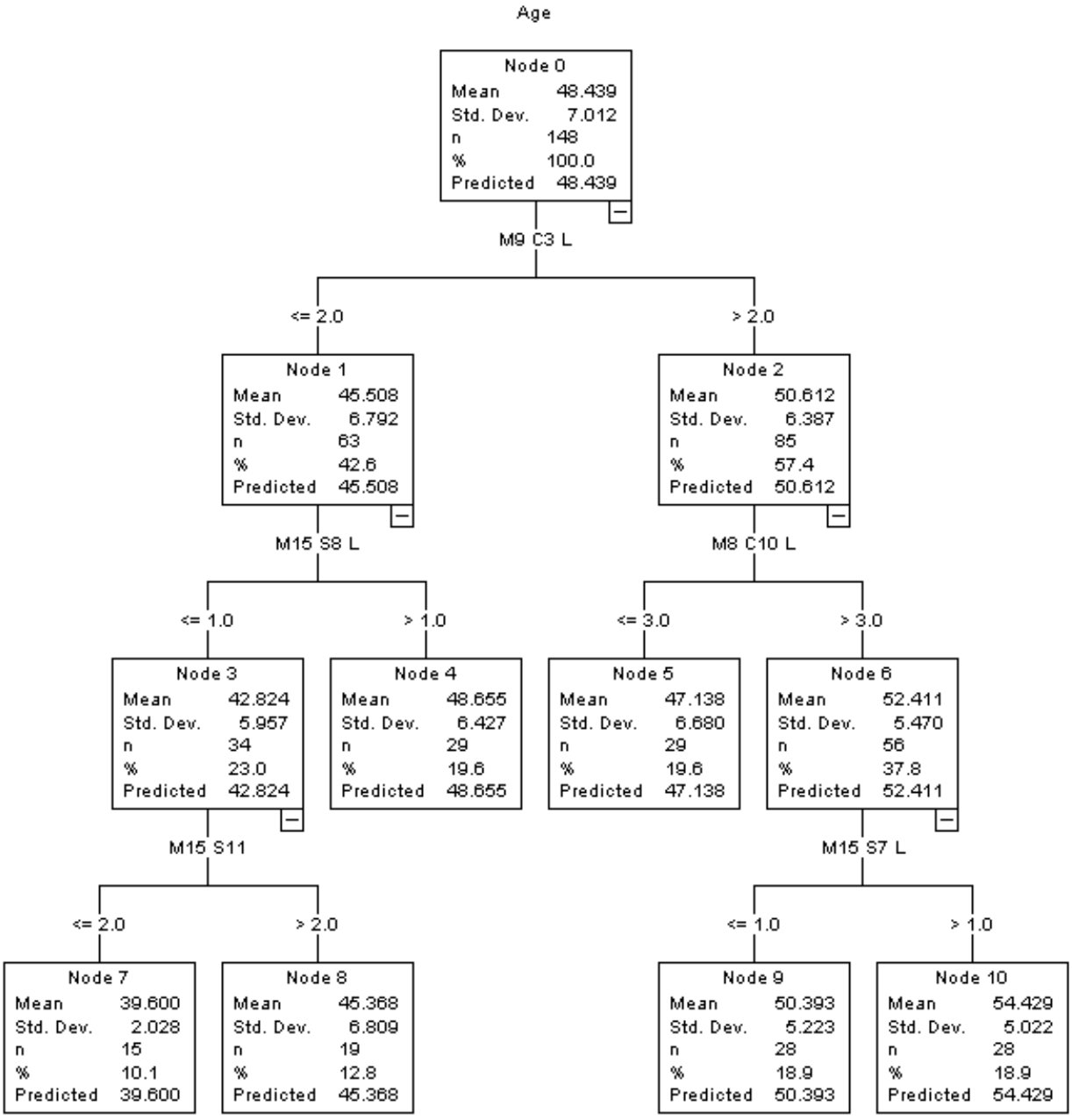


Figure C.9: CHAID 30/10 decision tree without age group 1 (23-35)

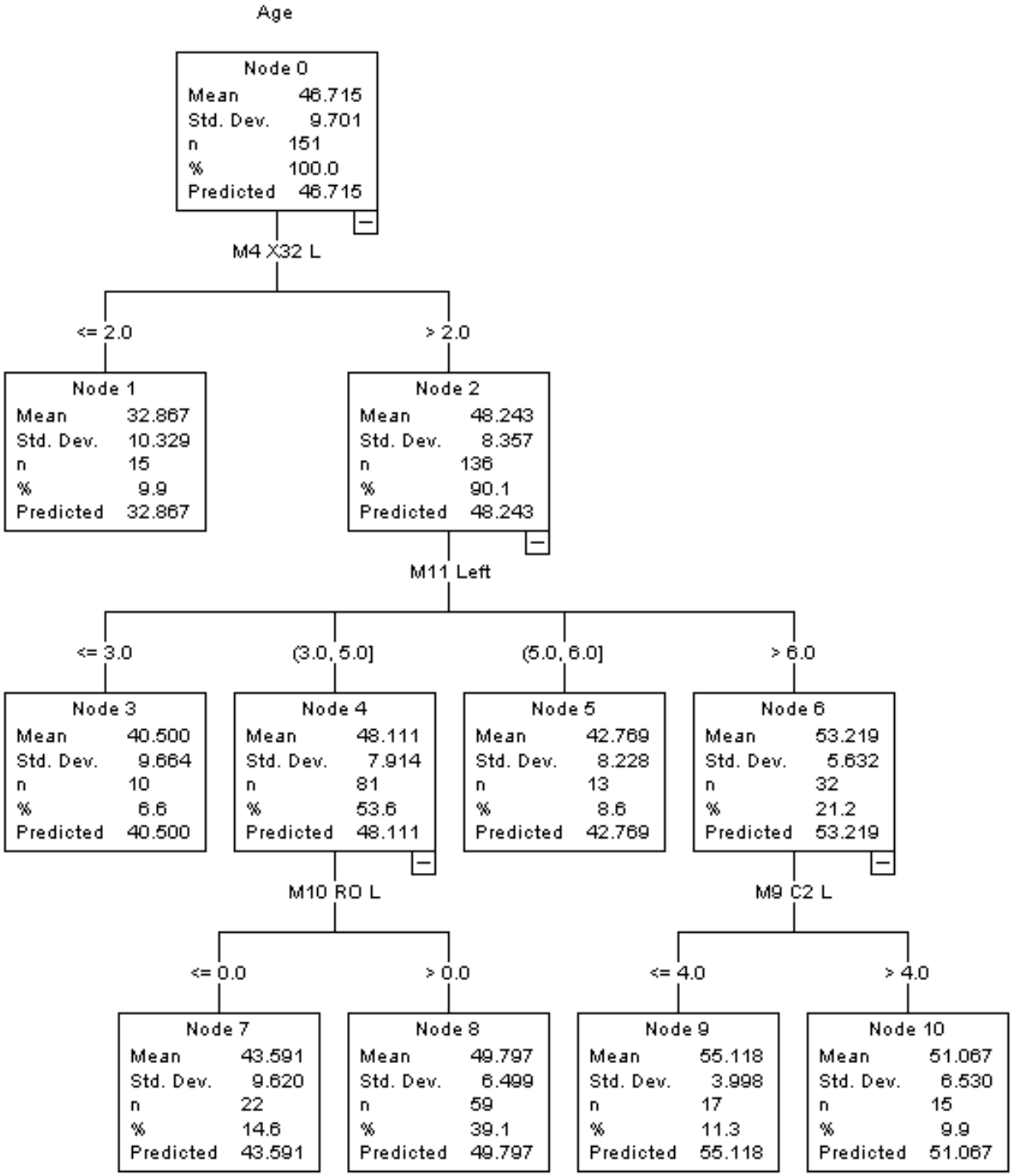


Figure C.10: CHAID 30/10 decision tree without age group 2 (36-40)

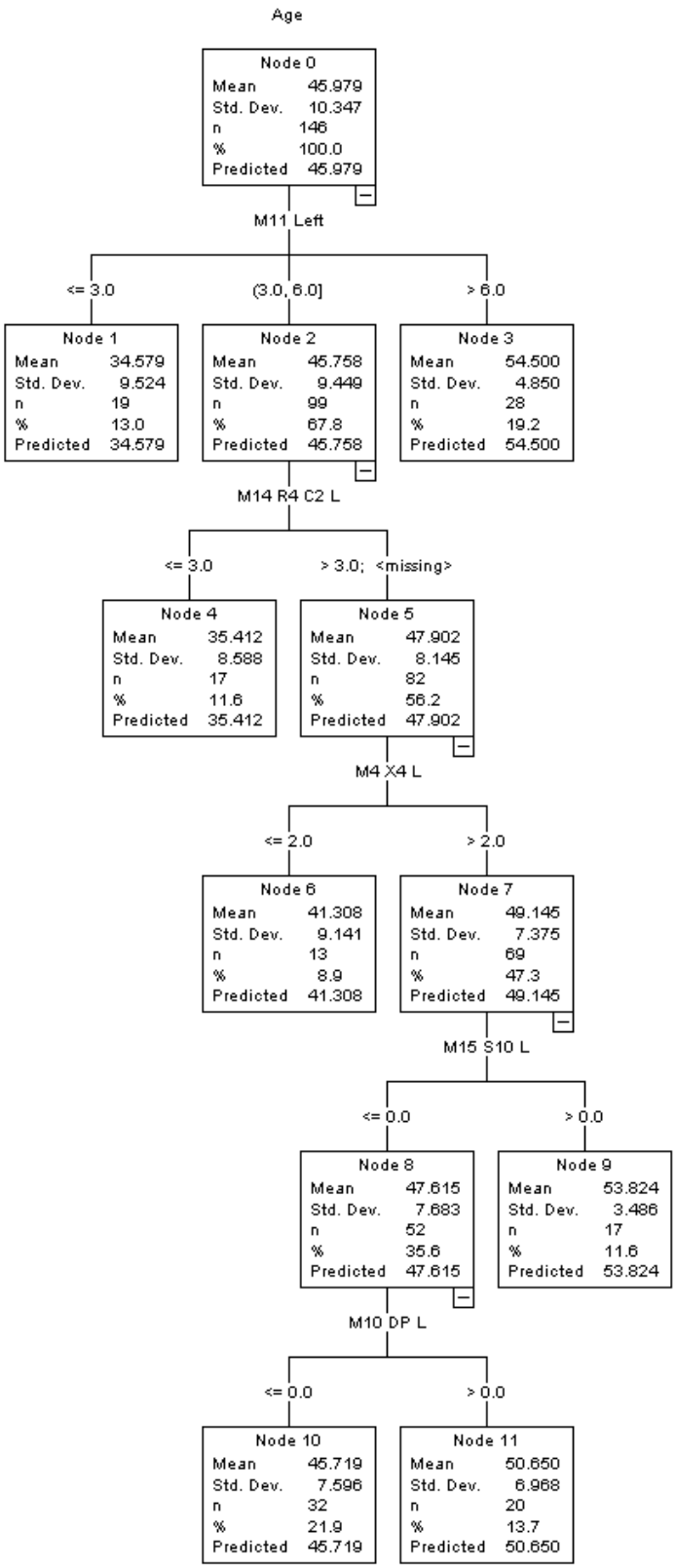


Figure C.11: CHAID 30/10 decision tree without age group 3 (41-45)

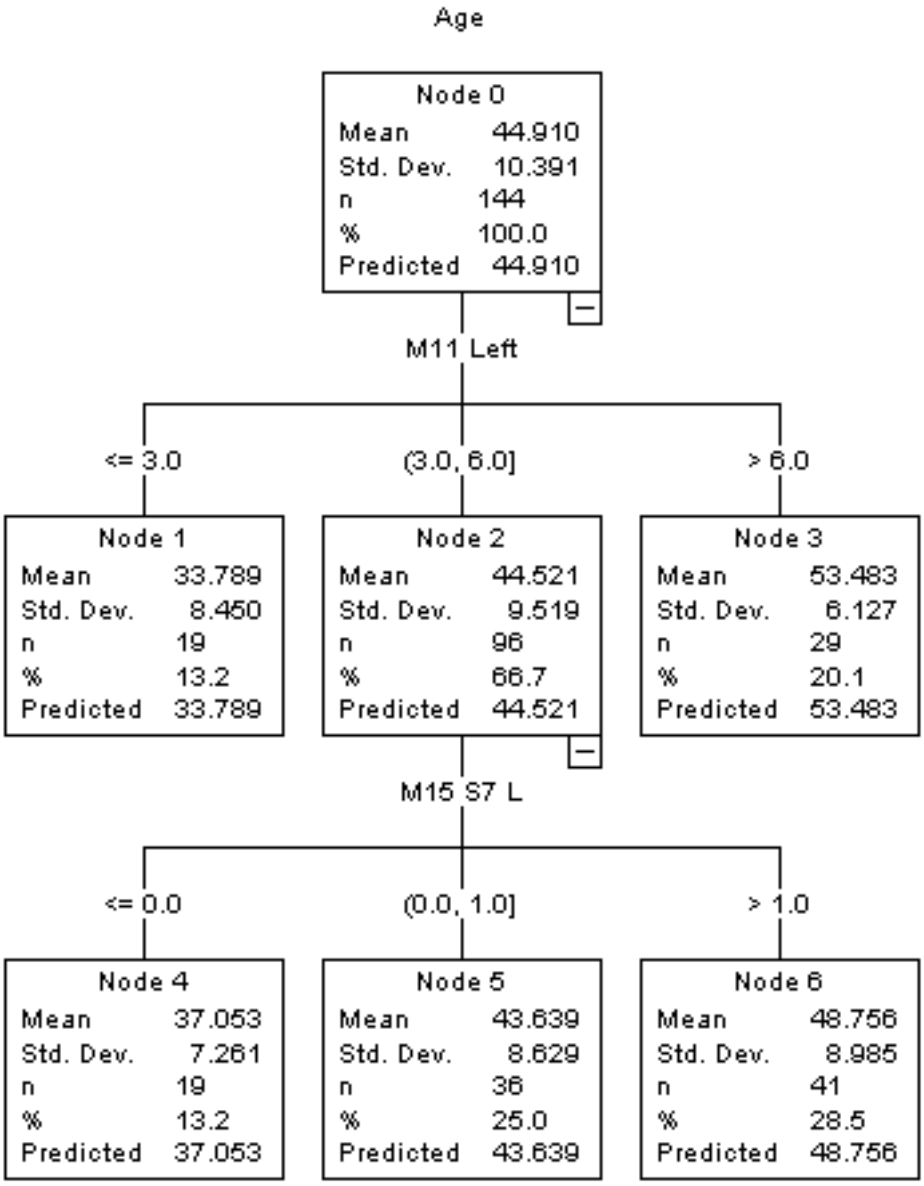


Figure C.12: CHAID 30/10 decision tree without age group 4 (46-50)

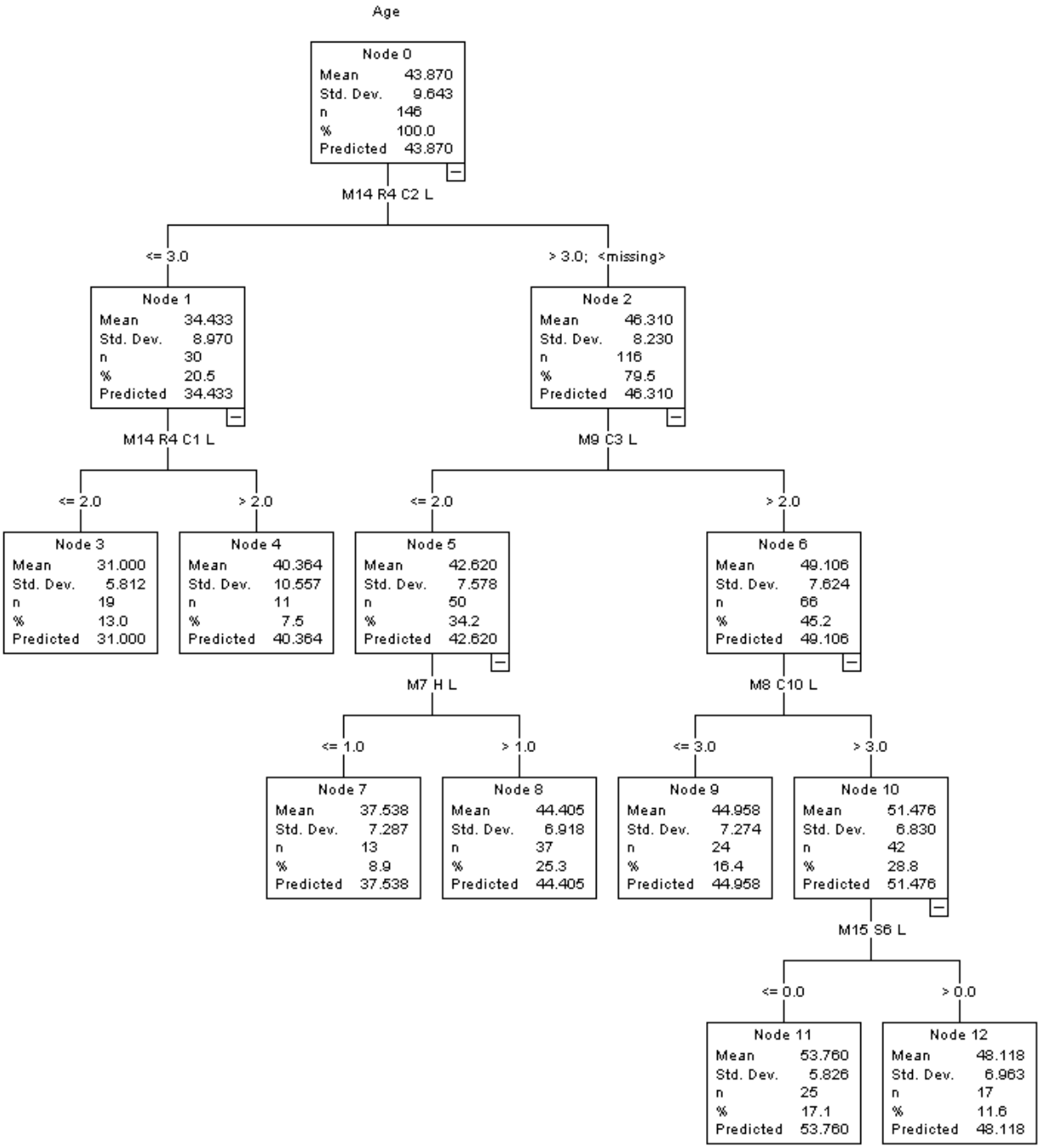


Figure C.13: CHAID 30/10 decision tree without age group 5 (51-55)

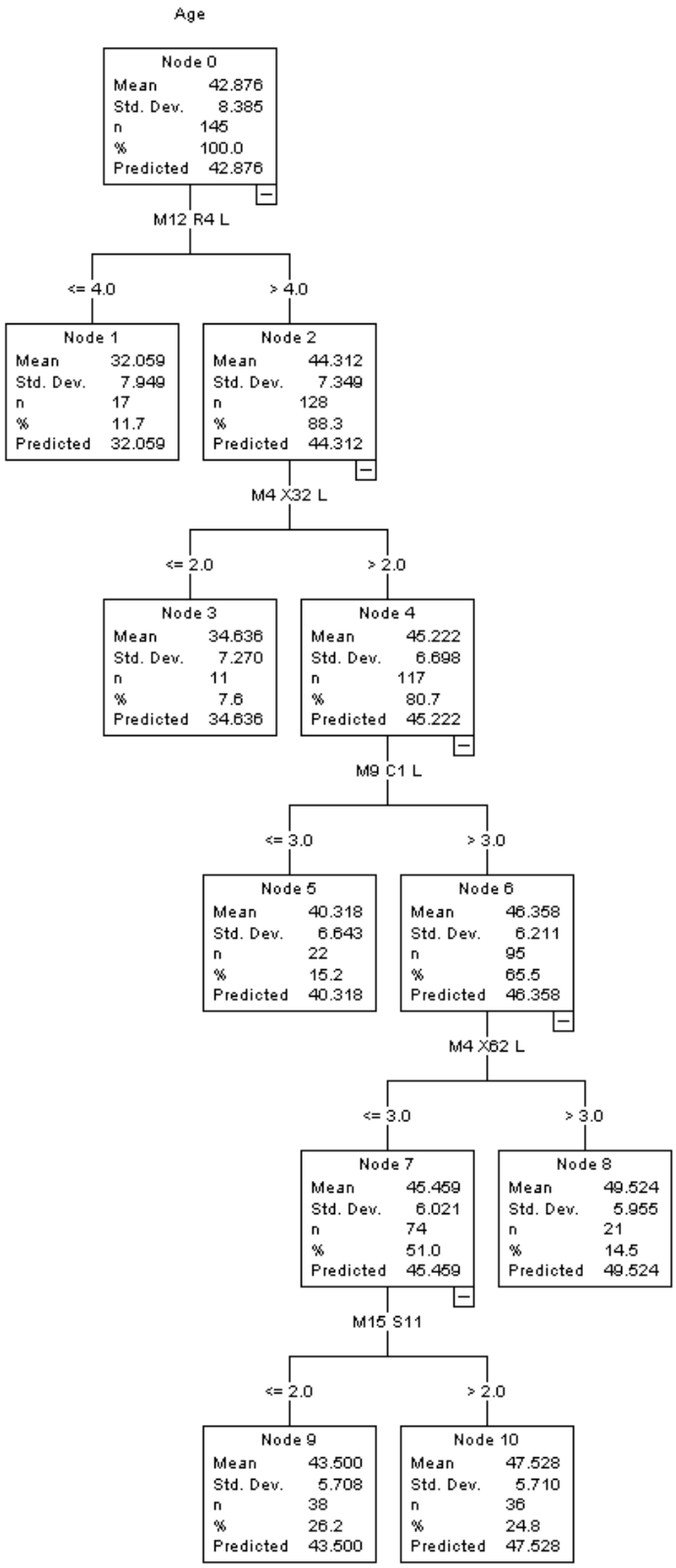


Figure C.14: CHAID 30/10 decision tree without age group 6 (56-60)

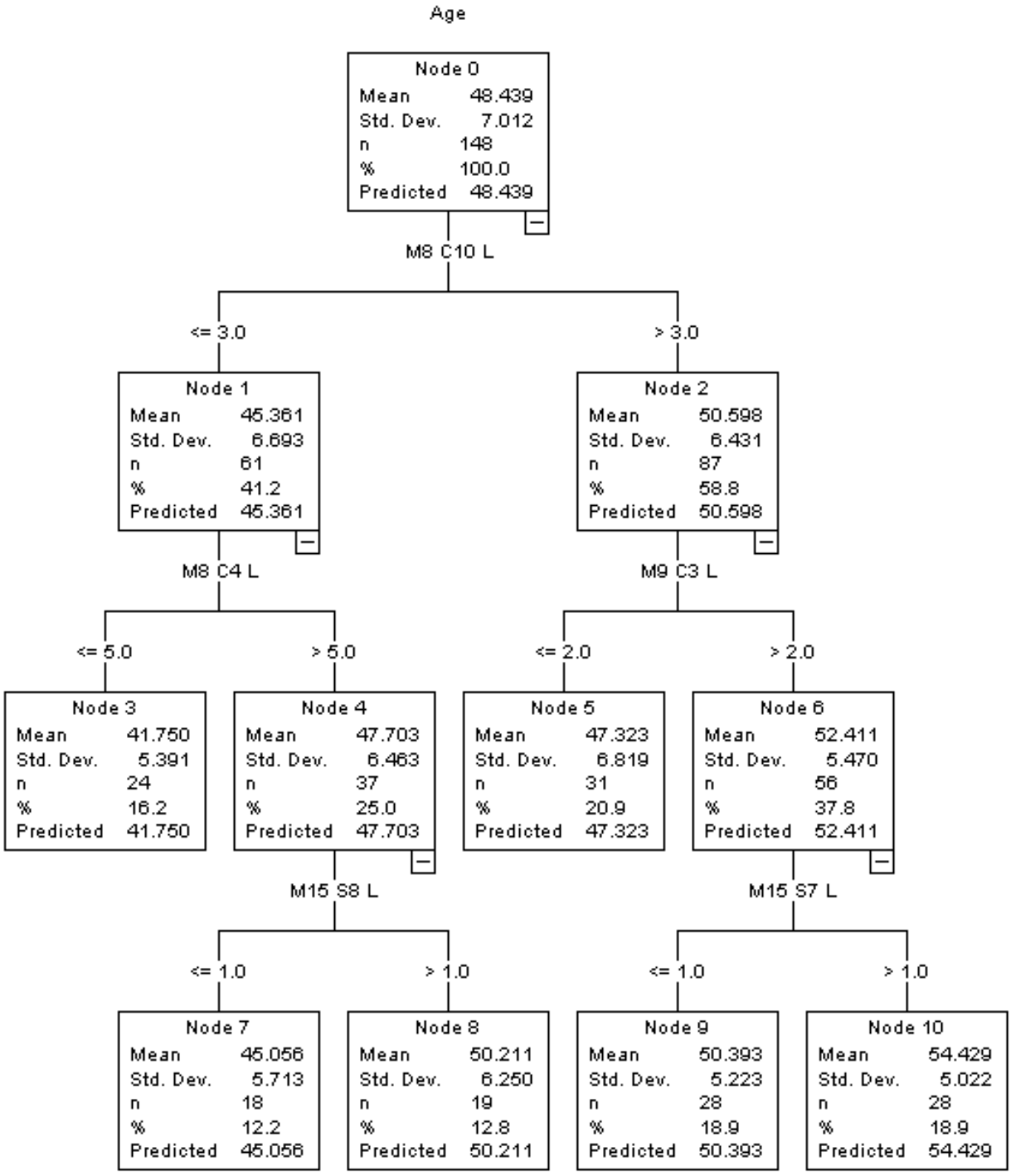


Figure C.15: CRT 30/10 decision tree without age group 1 (23-35)

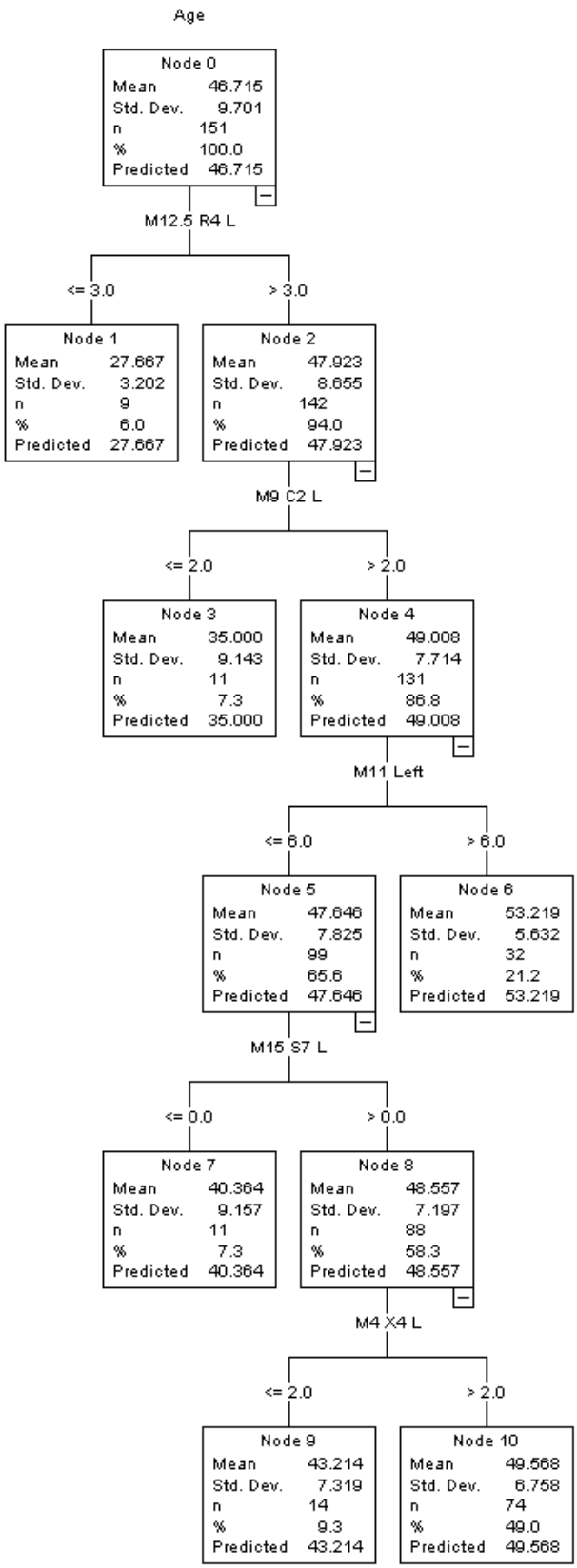


Figure C.16: CRT 30/8 decision tree without age group 2 (36-40)

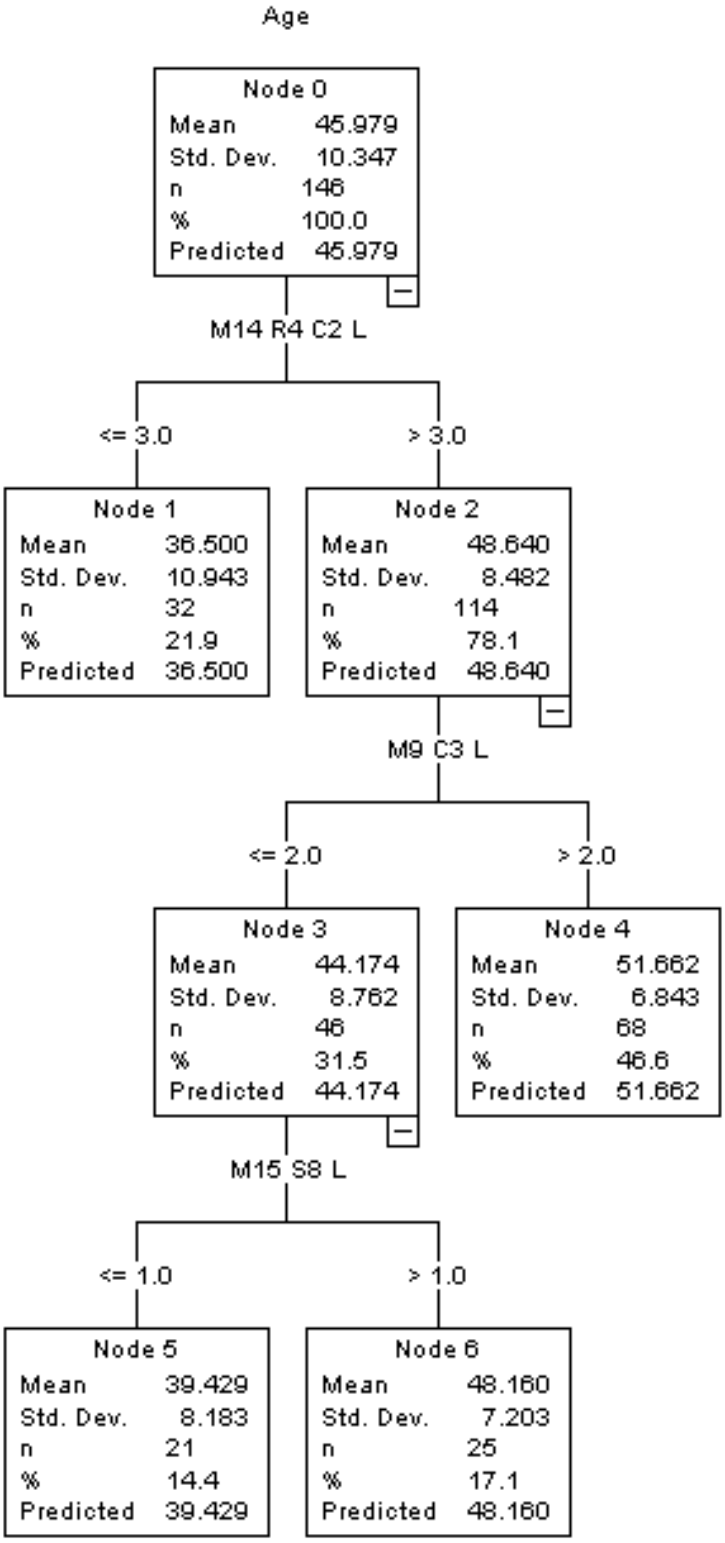


Figure C.17: CRT 30/10 decision tree without age group 3 (41-45)

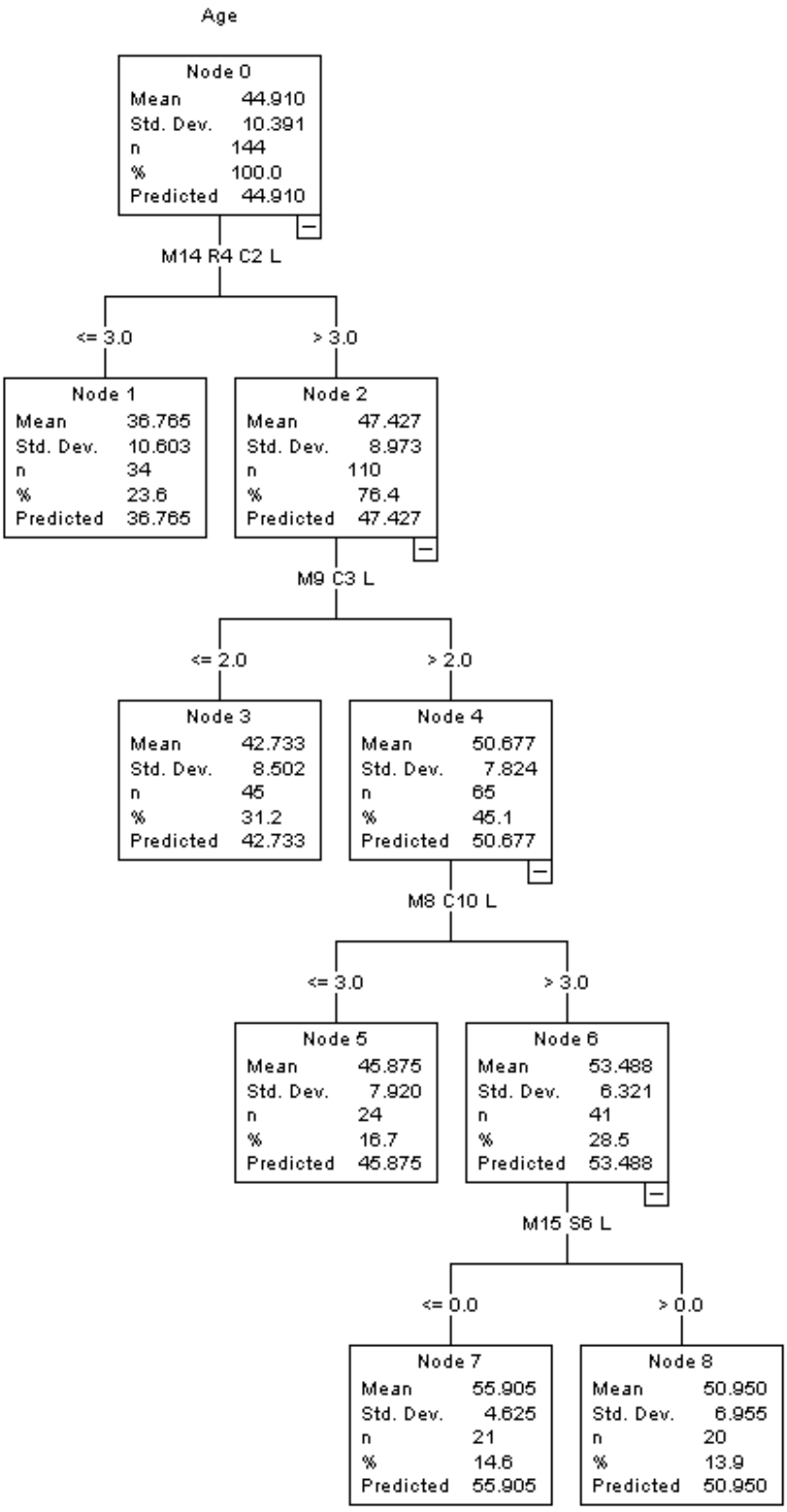


Figure C.18: CRT 30/10 decision tree without age group 4 (46-50)

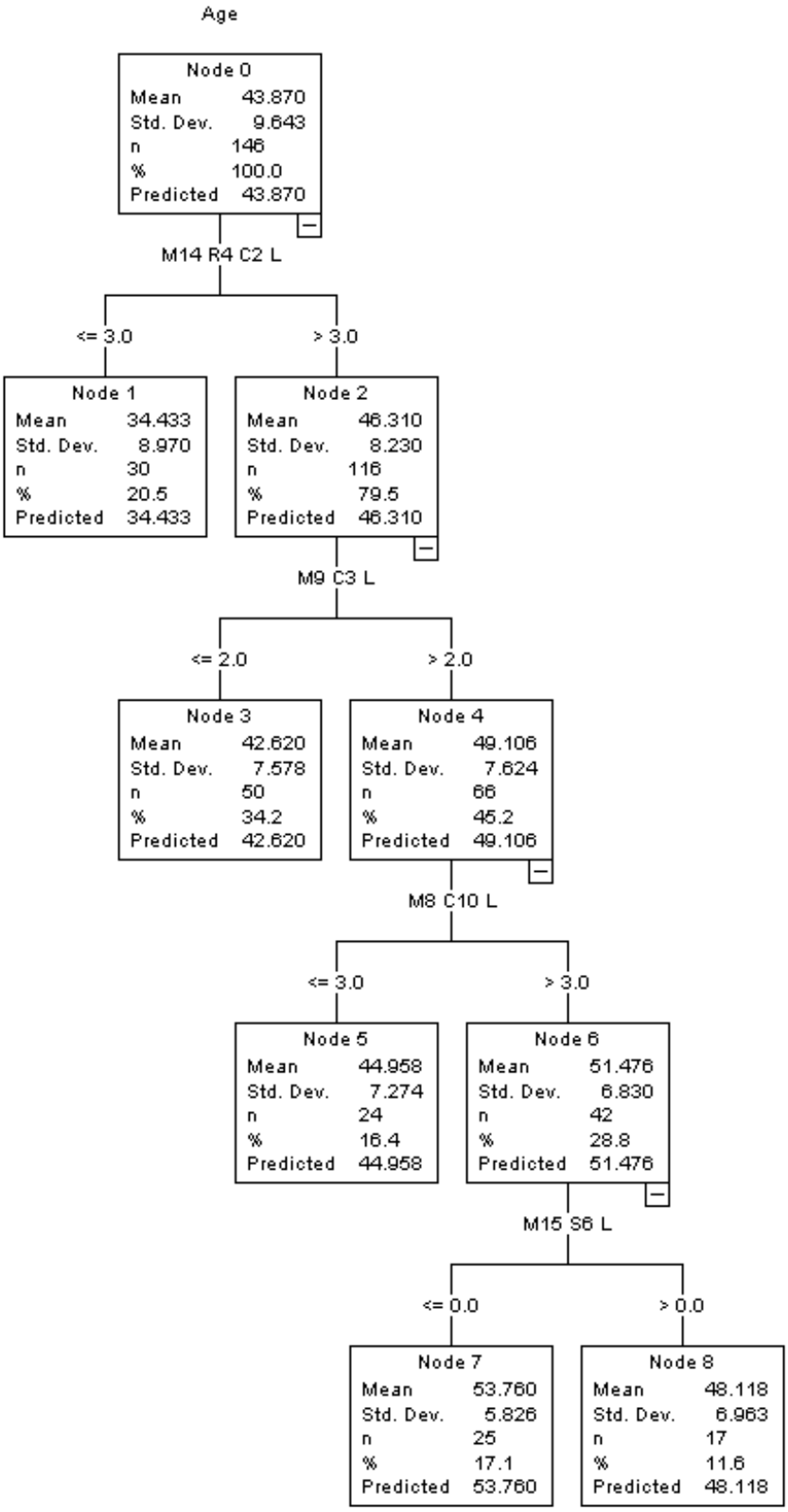


Figure C.19: CRT 30/10 decision tree without age group 5 (51-55)

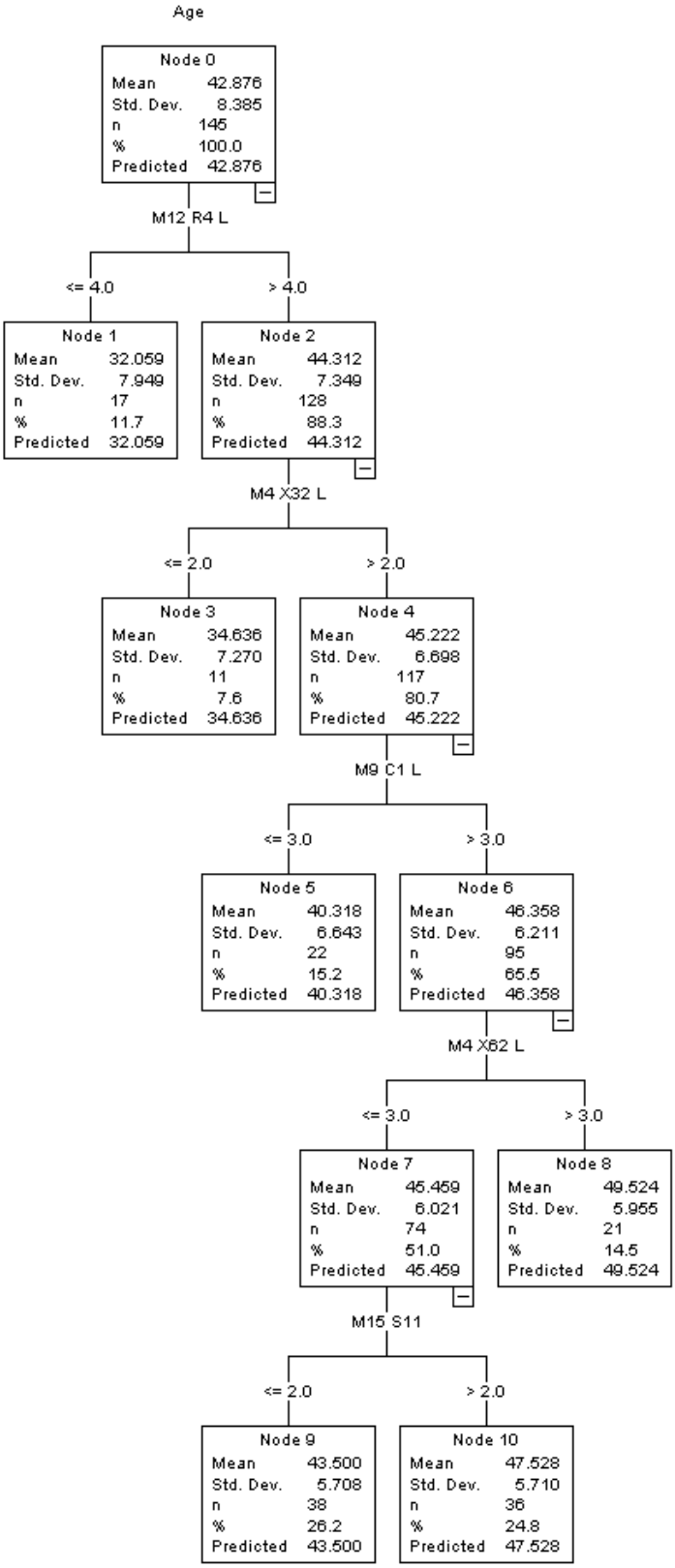


Figure C.20: CRT 30/10 decision tree without age group 6 (56-60)

Vita

Kevin Hufnagl has spent a lot of time studying throughout the United States. After graduating from High School in Iowa, he completed a double major in Mathematics and Archaeology from the University of Wisconsin--La Crosse. He continued by adding some physiological education at Montana State University and received his Masters of Science from Louisiana State University in Anthropology. He finally proceeded to the University of Tennessee in Knoxville, to finish out his education with an anthropological PhD. He is curious what the future holds for him and where he will experience it.