Pediatric Disparities in Sickle Cell Disease and Quantile Regression for Childhood Growth Charts with a Multi-institutional Databank of Electronic Health Records

Ryan Gallagher
BSc (University of Wisconsin - Eau Claire)

A capstone project in fulfillment
of the requirements for the degree of
Master of Arts in Biostatistics & Data Science

Division of Biostatistics
Institute for Health & Equity
Medical College of Wisconsin

Submitted December 5, 2023

Declaration

I hereby declare that this submission is my own work and that, to the best of my knowledge and belief, it contains no material previously published or written by another person nor material which to a substantial extent has been accepted for the award of any other degree or diploma of this institution or other institute of higher learning, except where due acknowledgement has been made in the text.

Ryan Gallagher

December 5, 2023

Abstract

This capstone project explores the implications of Sickle Cell Disease (SCD) on the growth patterns of pediatric patients, focusing on their Body Mass Index (BMI) trajectories. By utilizing electronic medical record (EMR) data from the TriNetX database, this research develops growth charts for children with SCD which include comparison against national BMI standards provided by the CDC. The study employs Quantile Regression, an advanced statistical method, to dissect the nuances and variability of BMI distributions across various quantiles, providing a granular perspective of growth trends in the SCD pediatric population.

The findings reveal significant disparities in BMI growth patterns between children with SCD and their healthy counterparts. This research offers evidence-based insights for healthcare providers who look to manage the distinct health needs of children with SCD. This project not only contributes to the growing body of knowledge on pediatric SCD and BMI but also underscores the potential of utilizing large clinical databases like TriNetX for in-depth health research.

Acknowledgements

I would like to express my appreciation to Dr. Rodney Sparapani, whose guidance shaped the direction and focus of this capstone project. I am also grateful to Dr. Ashima Singh for her contextual insight, helpful feedback and encouragement.

I extend my gratitude to the Division of Biostatistics for providing the necessary resources and environment conducive for research.

My sincere thanks go to my peers and classmates, for their camaraderie helped enrich my learning over the duration of this program.

I also wish to extend my heartfelt gratitude to my family and friends. Their constant encouragement and unwavering belief in my capabilities provided a foundation of support that was indispensable throughout this journey.

Contents

List of Figures

List of Tables

Nomenclature

List of Symbols

X	Predictor Variable
Y	Outcome
τ	Quantile
β_τ	Regression Coefficient Vector

List of Acronyms

MCW	Medical College of Wisconsin
SCD	Sickle Cell Disease
PI	Principal Investigator
CDC	Centers for Disease Control and Prevention
NHANES	National Health and Nutrition Examination Survey Data
NCHS	National Center for Health Statistics
EMR	Electronic Medical Record
BMI	Body Mass Index
ESRD	End Stage Renal Disease
CKD	Chronic Kidney Disease
LOINC	Logical Observation Identifiers Names and Codes
HbS	Hemoglobin S
HbSS	Sickle Cell Anemia
HbSC	Sickle Cell C Disease
HbSB	Sickle Cell Beta Thalassemia
HbAS	Sickle Cell Trait

Chapter 1

Background

1.1 Introduction

This project serves the purpose of fulfilling the Master's Capstone graduation requirement for the Masters of Arts degree in Biostatistics and Data Science at the Medical College of Wisconsin (MCW). Dr. Rodney Sparapani from the MCW Division of Biostatistics has served as my faculty mentor for this project. Dr. Ashima Singh from the MCW Department of Pediatrics serves as the principal investigator (PI).

1.2 Statement of Purpose

The purpose of this research is to analyze the the Body Mass Index (BMI) growth charts for children with Sickle Cell Disease (SCD). This investigation is a part of an ongoing initiative to leverage information from comprehensive electronic health record (EHR) repositories, built to make information available from a multitude of healthcare organizations, including hospitals, academic medical centers, and healthcare systems.

1.3 Research Objectives

This project seeks to examine existing, aggregated BMI data to form growth charts for children with sickle cell disease. We look to identify any significant deviations or differences in BMI growth patterns between children with sickle cell disease and the general pediatric population. Data for the general pediatric population exists in data repositories provided through the CDC.

Based on our findings, our goal is to provide evidence-based recommendations for healthcare practitioners and clinicians to improve the monitoring and management of nutritional health in pediatric patients with sickle cell disease.

1.4 Context

1.4.1 Epidemiology

Sickle cell disease is a hereditary condition where red blood cells, normally disc-shaped, become crescent or "sickle" shaped due to a genetic mutation affecting hemoglobin. This misshape can obstruct blood flow, causing serious complications like stroke, infections, and severe pain crises. Sickle cell disease predominantly affects individuals of African descent in the United States, with 1 in 13 Black or African American babies born with sickle cell trait, and 1 in every 365 born with the disease. Globally, it affects over 20 million individuals, with more than 100,000 cases in the United States alone (Kavanagh et al., 2022). Currently, one of the only definitive treatments for SCD is a bone marrow transplant (BMT). This procedure, which replaces the patient's marrow with healthy marrow from a donor, can effectively cure the disease. However, finding a suitable donor can be challenging, and the procedure carries significant risks, including rejection and serious infection (Walters et al., 1996).

1.4.2 Types of Sickle Cell Disease

Sickle Cell Disease is a collective term for a group of genetic disorders characterized by the production of abnormal hemoglobin, known as hemoglobin S (HbS). Among its various types, HbSS disease, commonly referred to as Sickle Cell Anemia, is the most prevalent and severe form. Another notable variant is HbSC disease, where individuals inherit one sickle cell gene and one gene for another abnormal hemoglobin called "C". This condition typically presents with milder symptoms compared to HbSS disease, but the severity can vary significantly among individuals. Patients with HbSC may experience some of the complications associated with sickle cell anemia, though generally to a lesser extent.

Apart from these, other forms include HbSβ-Thalassemia, which combines a sickle cell gene with a beta-thalassemia gene, and rarer variants like HbSD, HbSE, and HbSO. Sickle Cell Trait (HbAS) also falls under the SCD spectrum, where individuals carry one sickle cell gene but usually do not exhibit typical SCD symptoms.

1.4.3 Sickle Cell Disease & Malaria in Sub-Saharan Africa

Sickle Cell Disease has a profound impact on Sub-Saharan Africa, a region that bears a significant burden of this genetic disorder. This region has the highest prevalence of the disease globally, largely due to the genetic advantage that the sickle cell trait provides against malaria, a disease endemic in these areas. Individuals who are carriers of one sickle cell gene have a survival advantage in malaria-endemic areas, leading to the widespread presence of this gene in the population. However, when two carriers of the sickle cell trait have a child, there is a 25% chance with each pregnancy that the child will inherit SCD (Makani et al., 2013).

Malaria is caused by Plasmodium parasites, transmitted to humans through the bites of infected mosquitoes. This disease is highly dangerous, leading to severe symptoms like high fever, chills, and anemia, and is a major cause of death globally, especially in Sub-Saharan Africa. WHO (2023) estimates that malaria causes hundreds of

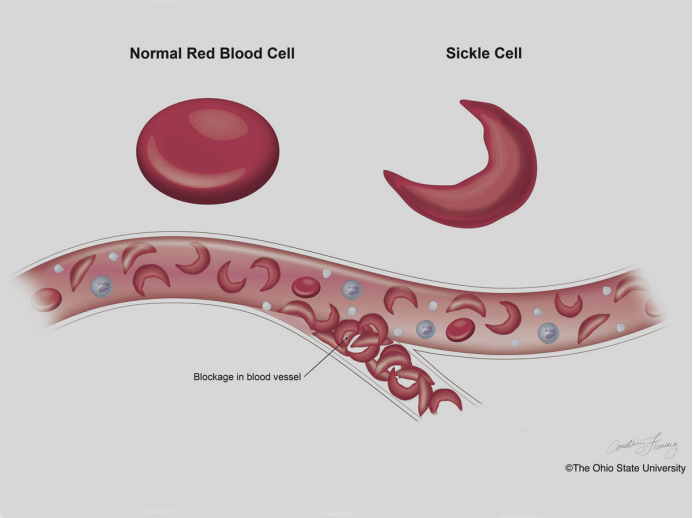

Figure 1.1 – Sickled Red Blood Cells, (Ohio State University, 2023)

thousands of deaths annually, predominantly among children. The sickle cell trait, involving one copy of the sickle cell gene, offers a protective advantage against malaria. This is due to the altered shape and reduced lifespan of red blood cells in carriers of the trait, making it more difficult for the malaria parasites to thrive and multiply. (Aidoo et al., 2002).

1.4.4 Pediatric Complications

Sickle cell disease presents a range of complications that can significantly impact the quality of life of affected individuals. Acute pain crises, triggered by sickled cells obstructing blood flow, manifest as sharp, intense pain throughout the body, often without warning. Chronic long-term pain is another concern, differing from crisis pain or pain stemming from organ damage, yet remains a common affliction. Individuals with SCD are also at a higher risk for nutrient and vitamin deficiencies, notably vitamins D, C, omega 3, and zinc, which can trigger crises or exacerbate existing complications. Delayed growth and puberty are consequences of the anemia associated with SCD, causing affected children to mature more slowly compared to their peers (CDC, 2023).

The effect of sickle cell disease on children are particularly complicated. According

to Children's Hospital of Philadelphia (2023), children are prone to a variety of complications that significantly impact their well-being. Aside from the expected anemic complications, children can be afflicted with ailments such as splenic sequestration which is a life-threatening situation where sickled cells get trapped in the spleen, requiring immediate medical attention. There's also Dactylitis or Hand-Foot Syndrome, often the first symptom of SCD in infants, presents as swelling and/or pain in hands or feet. Acute Chest Syndrome and stroke are severe complications caused by trapped sickle cells in the lungs and blockage of blood vessels to the brain, respectively. Children are also faced with the potential for learning difficulties and other cognitive impairments due to brain damage.

SCD profoundly influences not only the physical health but also the psychosocial well-being of affected individuals. Children with SCD often face significant psychosocial challenges, including recurring pain crises that disrupt daily activities, leading to frequent school absences and social isolation. This disruption can result in considerable psychological stress, heightened anxiety, and depression, further complicating their cognitive and emotional development (Brown et al., 2016). Studies, such as those by DeBaun and Telfair (2012), have highlighted the impact of SCD on mental health and educational outcomes, noting that children with SCD may experience attention deficits and learning difficulties. Additionally, the extensive impact of SCD on various organ systems is a crucial aspect of its complexity. The disease is known to affect multiple organs, leading to complications such as increased stroke risk, pulmonary hypertension, renal impairment. Rees et al. (2010) underscores that these complications arise from the chronic hemolytic anemia and vaso-occlusion characteristic of SCD, necessitating a multidisciplinary approach for comprehensive care.

1.4.5 Newborn Screening

In the United States, newborn screening is a public health program aimed at the early identification of conditions that can affect a child's long-term health or survival. This process typically occurs within the first few days of a newborn's life. It involves a

series of tests conducted on newborns to screen for genetic, endocrinologic, metabolic, and hematologic diseases. The primary method is a blood test, where a few drops of blood are taken from a newborn's heel. This blood is then analyzed for a panel of conditions, the specifics of which can vary by state or country (AAFP, 2008).

Sickle Cell Disease is one of the critical conditions screened for in newborns. Given the serious health challenges and complications associated with SCD, early identification is vital. With early detection through newborn screening, children with SCD can receive prompt and appropriate medical care. This can include interventions like vaccinations and antibiotics to prevent infections, nutritional management, and regular health check-ups to monitor and manage the disease effectively.

The initiation of early treatment and routine follow-up care significantly improves the survival and quality of life of these children. Furthermore, newborn screening for SCD plays a crucial role in educating and preparing parents about the condition, its implications, and the need for continuous care and monitoring. It also aids in reducing the incidence of related complications such as stroke, pain episodes, and organ damage, thereby mitigating the overall burden of the disease on both the affected individuals and the healthcare system (Upadhye et al., 2016).

1.5 Theoretical Framework

1.5.1 BMI

Body Mass Index (BMI) is a calculation derived from an individual's height and weight. It's calculated as:

$$\text{BMI} = \frac{\text{weight (kg)}}{\text{height (m)}^2} \tag{1.1}$$

For children and teens, it's age- and sex-specific, known as BMI-for-age. Using growth charts (CDC, 2023), a child's BMI is plotted to determine its percentile, which classi-

fies their weight status. These categories include underweight, healthy weight, overweight, and obese. Growth charts are segmented for male and female children due to inherent differences in biological and physiological growth patterns, where these variations become particularly evident during puberty when boys and girls experience different growth speeds and developmental changes.

According to the AHA (2023), a child's BMI is an indicator that can provide insights into their overall health. A significantly low BMI in children can be indicative of potential health concerns. While the specific cause can vary, it often signals issues such as inadequate nutrition, which may impede proper growth and development. Conversely, a high BMI in children is commonly associated with excess body weight. This condition can predispose them to a range of immediate and long-term health issues. These include an increased risk of developing chronic conditions such as type 2 diabetes and cardiovascular problems. Additionally, high BMI in children has been linked to psychosocial issues, including low self-esteem, further underscoring the need for careful health monitoring and intervention when necessary.

1.5.2 Growth Charts

Growth charts are essential tools in pediatric healthcare, used to monitor and assess the growth and development of children. The primary purpose of growth charts is to track a child's growth over time, comparing it with standardized growth patterns for children of the same age and sex. Implementation of growth charts typically involves regularly measuring a child's height, weight, and sometimes head circumference, and plotting these measurements on the chart. These charts offer a visual representation of a child's growth trajectory compared to the established norms. These charts include specific percentiles that indicate the relative position of a child's growth measure among their peers (CDC, 2023).

Growth charts serve several functions. Firstly, they help in identifying children who are growing at an abnormal rate, either too fast or too slow. For example, deviations from standard growth patterns may indicate conditions like malnutrition, obesity,

growth hormone deficiencies, or genetic disorders. Secondly, they are useful for tracking the progress of children with chronic health conditions such as SCD, assessing how well these conditions are being managed. Thirdly, growth charts can be instrumental in reassuring parents and caregivers about their child's healthy development or in identifying when further evaluation and intervention might be necessary (AHA, 2023).

1.5.3 LMS Method vs. Quantile Regression

While this paper will be creating growth charts using quantile regression, the standard for developing growth charts is the LMS method developed by Cole and Green (1992). The LMS method is different in that it creates smoothed percentile curves that represent the distribution of a particular growth measurement across a population. The method uses three parameters – L (skewness), M (median), and S (coefficient of variation) – to describe this distribution. By using the L parameter for power transformation, the LMS method adjusts the data so that it aligns more closely with a normal distribution. The M and S are for adjusting the percentile curves to accurately reflect the natural variation and skewness in the growth data. This ensures that the growth charts are representative of the actual population and can cater to diverse growth patterns.

The LMS method and Quantile Regression, though both used for analyzing data distributions, are distinct in their applications and characteristics. The LMS method focuses on summarizing the distribution of measurements in relation to a covariate, using three parameters to transform skewed data into a normal distribution. In contrast, Quantile Regression provides a broader analysis of the relationship between variables across their entire distribution. It fits multiple lines for different quantiles, without transformation. Unlike LMS, Quantile Regression does not assume a normal distribution of residuals and is less sensitive to outliers, making it more adaptable to various types of real-world data.

1.6 Literature Review

Hall et al. (2018) assessed the prevalence of high BMI in children with SCD and its correlation with disease severity. The study included 385 patients aged 2-18 years with different forms of SCD. Disease severity was measured by parameters including hospital admission rates, lactate dehydrogenase levels, and obstructive sleep apnea. The results showed that 17% of the children were overweight or obese, but high BMI did not correlate with disease severity in this cohort. Notably, obesity was more prevalent among children with HbSC (a form of SCD) and females. The study concluded that further prospective studies are needed to determine the long-term effects of BMI on disease severity and outcome in SCD patients

Jackson et al. (2022) suggests that there might be a correlation between BMI and hemoglobin levels in children with SCD. Specifically, this study finds higher hemoglobin values in children who were overweight or obese compared to those with normal BMI. This is particularly interesting when you consider that Mpalampa et al. (2012) found that the levels of fetal hemoglobin (HBF) are generally found to be inversely proportional to the severity of SCD. This suggests that a BMI measurement might be useful in predicting SCD severity - which could be important in the clinical course of children with sickle cell disease.

Chapter 2

Methods

2.1 Quantile Regression

2.1.1 Overview

Quantile regression, introduced by Koenker and Bassett (1978), represents an extension to the classical linear regression model. Unlike traditional regression methods that primarily focus on estimating the mean of the dependent variable based on independent variables, quantile regression explores the estimation of various quantiles of the dependent variable. This approach provides a more detailed perspective on the relationship between variables.

A quantile is a measure which divides a frequency distribution into equal groups, each containing the same fraction of the total population. In essence, it segments the data based on each data point's position in the overall distribution. For example, the median divides the data so that 50% of the values lie below it. Other common quantiles include quartiles, which divide the data into four equal parts, and percentiles, which divide it into 100 equal parts. These quantiles are useful in allowing for an examination of how different parts of the distribution of the dependent variable relate to the independent variables.

In quantile regression, we gain understanding into how the predictors affect not just the median (or the 0.5 quantile) of the response variable, but also other quantiles like the 0.10 or 0.95 quantile. This method is particularly useful when the relationships between variables vary across different points of the distribution or when the response variable's distribution is skewed or contains outliers. By focusing on specific quantiles instead of the overall mean, quantile regression offers a more robust analysis.

Quantile regression is especially beneficial in situations with non-uniform variance (heteroscedasticity) within the data, providing a more nuanced understanding of the underlying relationships. This is because (unlike OLS) quantile regression does not make assumptions about the variance of the error terms. It estimates the conditional median (or other quantiles/percentiles) of the dependent variable, making it less sensitive to variations across the range of predictors (Machado and Silva, 2013).

Quantile regression can be implemented using a variety of statistical software:

- **R**: '`quantreg`', developed by Roger Koenker, includes functions for fitting, conducting inference, and visualizing quantile regression models.
- **SAS**: '`PROC QUANTREG`' performs quantile regression analysis with options for fitting and diagnosing quantile regression models. The SAS documentation also provides detailed examples on visualizing these models.
- **Python**: '`statsmodels`' is a library which offers functions for quantile regression that integreates will with other Python data analysis libraries.

2.1.2 Application

Quantile regression is conducted by focusing on the estimation of conditional quantiles of the response variable, providing a way to understand how these quantiles vary with changes in predictor variables. This methodology is notably different from ordinary least squares (OLS) regression, which centers on estimating the mean or average effect of the predictors.

Consider a response variable Y and a matrix of predictor variables X. In quantile regression, we aim to estimate the conditional quantile, or the τ-th quantile (where $0 < \tau < 1$), of Y given X. This is expressed as $Q_Y(\tau|X)$, where $Q_Y(\tau|X)$ denotes the τ-th quantile of the response variable Y given predictors X and β_τ. β_τ represents the vector coefficients to be estimated for the τ-th quantile:

$$Q_Y(\tau|X) = X\beta_\tau + \epsilon_\tau$$

where ϵ_τ represents the error term for the τ-th quantile. This term captures the the deviation of the actual value of the dependent variable from the value predicted at the τ-th quantile. This error term is analogous to what is defined in OLS, but specific to the quantile being estimated.

To find the quantile estimates, β_τ, quantile regression minimizes the sum of weighted absolute residuals, where the weights depend on the quantile τ being estimated. The function for estimating the τ-th quantile estimates is known as the *loss function*:

$$\rho(\tau, u_i) = \begin{cases} \tau u_i & \text{if } u_i \geq 0 \\ (\tau - 1)u_i & \text{if } u_i < 0 \end{cases} \tag{2.1}$$

where u_i is the residual error term for an individual observation, otherwise known as the difference between the observed value of the dependent variable and its predicted value from the model: $u_i = y_i - x_i^t\beta_\tau$. In the loss function, when the predicted value is less than the actual value (i.e. positive residual), the residual is multiplied by τ. This means if we are looking at a high quantile ($\tau = 0.90$), the function heavily penalized under-predictions because these residuals contribute more to the loss. Conversely, when the predicted value exceeds the actual value (i.e. negative residuals), the residual is multiplied by $\tau - 1$. In this case, over predictions are penalized more when τ is a lower quantile (such as $\tau = 0.1$). We can see how when $\tau = 0.5$ (the median) that the loss function treats overestimations and underestimations symmetrically, though asymmetry skews this penalty either towards overestimation or underestimation depending on whether τ is greater or less than 0.5.

As mentioned, the objective in quantile regression is to find β_τ by minimizing the sum of the loss function $\rho_\tau(u)$ across all observations. This is accomplished by solving the minimization problem:

$$\min \sum_{i=1}^{n} \rho(\tau, y_i - x_i^t \beta_\tau)$$

where the coefficients are interpreted as: the change in the specified quantile of the dependent variable for a one-unit change in the independent variable.

2.1.3 Use Cases

Quantile regression finds applications across a variety of academic disciplines. In economics and finance, it's used to analyze various aspects such as earnings inequality, financial returns, and risk management. In Fournier and Koske (2012), quantile regression is used to analyze the determinants of labor earnings across different segments of the income distribution. This method helps in understanding how various factors, such as education and employment contracts, impact earnings in different countries.

In ecology, Cade and Noon (2003) report on an analysis of Lahontan cutthroat trout abundance in relation to stream width-to-depth ratio. Here, quantile regression revealed a nonlinear negative relationship at higher percentiles of trout densities, a detail that would have been missed using mean regression estimates alone. In their results, the authors note how quantile regression can be used to estimate changes along the upper boundaries of conditional distributions, which is often of significant interest in ecological studies.

Application of quantile regression is also found in medical research. An example can be found in a case study examining gender differences in the timeliness of thrombolytic therapy for patients with acute myocardial infarction. Authors, Austin et al. (2005), used quantile regression to allow for a more comprehensive assessment of how different quantiles of treatment delays change with patient characteristics. The study

found that females were more likely to experience delays in thrombolytic treatment compared to males, and that gender had a greater impact on those patients who experienced the longest delays in treatment.

2.1.4 Extensions and Recent Developments

Extensions of quantile regression have been developed in the pursuit of robustness, flexibility, and accuracy. Popular extensions of quantile regression methods are as follows.

- **Quantile Regression Forests (QRF)**: Quantile Regression Forests, an extension of Random Forests, are used for estimating conditional quantiles of a response variable. Developed by Meinshausen and Ridgeway (2006), QRF is useful in cases where the relationship between the independent and dependent variables is non-linear or complex. QRF gives a non-parametric and accurate way of estimating conditional quantiles for high-dimensional predictor variable.

- **Quantile Regression Neural Networks (QRNN)**: Quantile Regression Neural Networks are, aptly named, the extension of quantile regression to the framework of neural networks. QRNN is an appropriate approach in situations where the conditional distribution of the response variable is skewed. QRNN finds application in financial forecasting and weather prediction (Yang et al., 2013).

- **Smooth Additive Quantile Regression Models (QGAM)**: Smooth Additive Quantile Regression Models are additive models which allow for smooth estimation of quantiles for the response distribution. QGAM are flexible models which accomodate non-linear relationships through additive smooth functions. They find application in environmental modeling, epidemiology, and economics (Stasinopoulos and Rigby, 2008).

2.2 Data Extraction

2.2.1 TriNetX

This project is a part of an ongoing initiative to leverage the TriNetX database to conduct observational, pediatric SCD research. TriNetX is a mainly US multi-national health research network platform that houses real-world patient data from sources like electronic health records and insurance claims. By connecting healthcare organizations, pharmaceutical firms, and contract research organizations, TriNetX fosters collaborative research while offering an extensive array of de-identified patient data. TriNetX does this while ensuring data privacy and compliance.

Within TriNetX is a querying tool which allows access to it's conglomerate database. From this, we are able to extract over 90,000 patients with a diagnosis code for SCD from more than 30 national healthcare organizations between 2010-2020. After filtering against improper coding, we expect to account for about 25% of the estimated cases in the United States. The codes used for initial extraction were the following: D57.0 (Hb-SS disease with crisis), D57.1 (Sickle-cell disease without crisis), D57.2 (Sickle-cell/Hb-C disease), D57.4 (Sickle-cell thalassemia), and D57.8 (Other sickle-cell disorders).

Data retrieved from TriNetX is formatted across multiple '.csv' files. There are 20 data files in total. Each included table is linked by *Patient ID* and/or *Encounter ID* The most relevant files are summarized as follows:

- **Diagnosis Table**: Contains information on patient diagnoses. It includes the diagnosis code, the code system (like ICD-9 or ICD-10), indicators for principal diagnosis, admitting diagnosis, and the patient's reason for visit. Each diagnosis is date-stamped.
- **Encounter Table**: This table includes details about healthcare encounters. The table records the start and end dates of encounters, the type of care setting (like Ambulatory, Inpatient, etc.), and flags to indicate whether dates or the

encounter itself was derived by TriNetX. The data source can be TriNetX or EMR.

- **Lab Result Table**: This table records lab test results for patients. It includes the lab test code and code system (like LOINC), the numeric and text results, units of measure, and the date of the test.
- **Patient Demographic Table**: Provides demographic information for each patient, including a unique ID, sex, race, ethnicity, marital status, year of birth, and geographical location. It also records if the patient is deceased, along with the death date source ID.
- **Procedure Table**: Contains details on medical procedures performed. It includes the procedure code, code system (like CPT or HCPCS), and the procedure date.
- **Vital Signs Table**: Tracks vital sign measurements. It lists the code and code system for each vital sign, the recorded value (numeric and text), units of measure, and the date of recording.

These tables in the TriNetX healthcare database collectively offer a holistic view of a patient's medical history. This integration allows for a comprehensive understanding of a patient's health status, treatment history, and healthcare interactions, facilitating more informed clinical decision-making and enabling detailed health research.

2.2.2 NHANES/CDC Data

The National Health and Nutrition Examination Survey (NHANES) is a program of studies designed to assess the health and nutritional status of adults and children in the United States. Conducted by the National Center for Health Statistics (NCHS), part of the Centers for Disease Control and Prevention (CDC), NHANES combines interviews and physical examinations to gather comprehensive data on various health aspects, including BMI (Centers for Disease Control and Prevention, 2023).

The data provided by NHANES is a comprehensive dataset that contains BMI percentile data, stratified by age and sex. This dataset is formatted to provide a view of BMI distribution across various age groups, measured in months, for both sexes. Each row corresponds to a specific age, starting from 24 months, and includes percentile values such as P3, P5, P10, P25, P50 (median), P75, P85, P90, P95, and P97. These percentiles are chosen to evaluate a child's BMI in relation to standardized growth patterns.

2.3 Inclusion/Exclusion Criteria

Patients were selected into our study based on satisfaction of a mix of criteria which ensured accuracy of diagnosis and availability of follow-up data:

1. 90,000 patients were gathered from TriNetX with diagnosis codes for SCD.
2. Patients were included whose encounter date occurred between 2016 and 2019 to favor accuracy of diagnosis.This range begins at the advent of ICD-10 while ending before the coding complications introduced with the COVID-19 pandemic.
3. Patients were excluded who had reported death dates outside of our date range.
4. Patients where excluded who reported having Sickle Cell Trait.
5. Patients were excluded if they did not have at least three occurrences of our ICD-10 SCD codes from our expanded list. This once again favored accuracy of diagnosis.
6. Patients were excluded based on age criteria, starting with the exclusion of individuals born before 1997, the earliest acceptable birth year for our study.
7. Patients were excluded who turned 18 during our follow-up period as they no longer met our definition of children.

8. Patients failing to meet our follow-up time criterion, which necessitated a minimum span of 365 days between the earliest and final recorded encounter dates were excluded.

9. Patients were excluded if they did not have two consecutive years of follow-up, with a provision for a one-year gap between these two years when required

10. Patients were included if they had available BMI data.

After exclusions were applied to our initial 90,000 individuals, we were left with 4559 children confidently identified as having SCD within our selected date range. Of these, we found that 1220 have BMI data available according to the LOINC codes found in the vitals dataset. Among these, 632 are male and 588 are female.

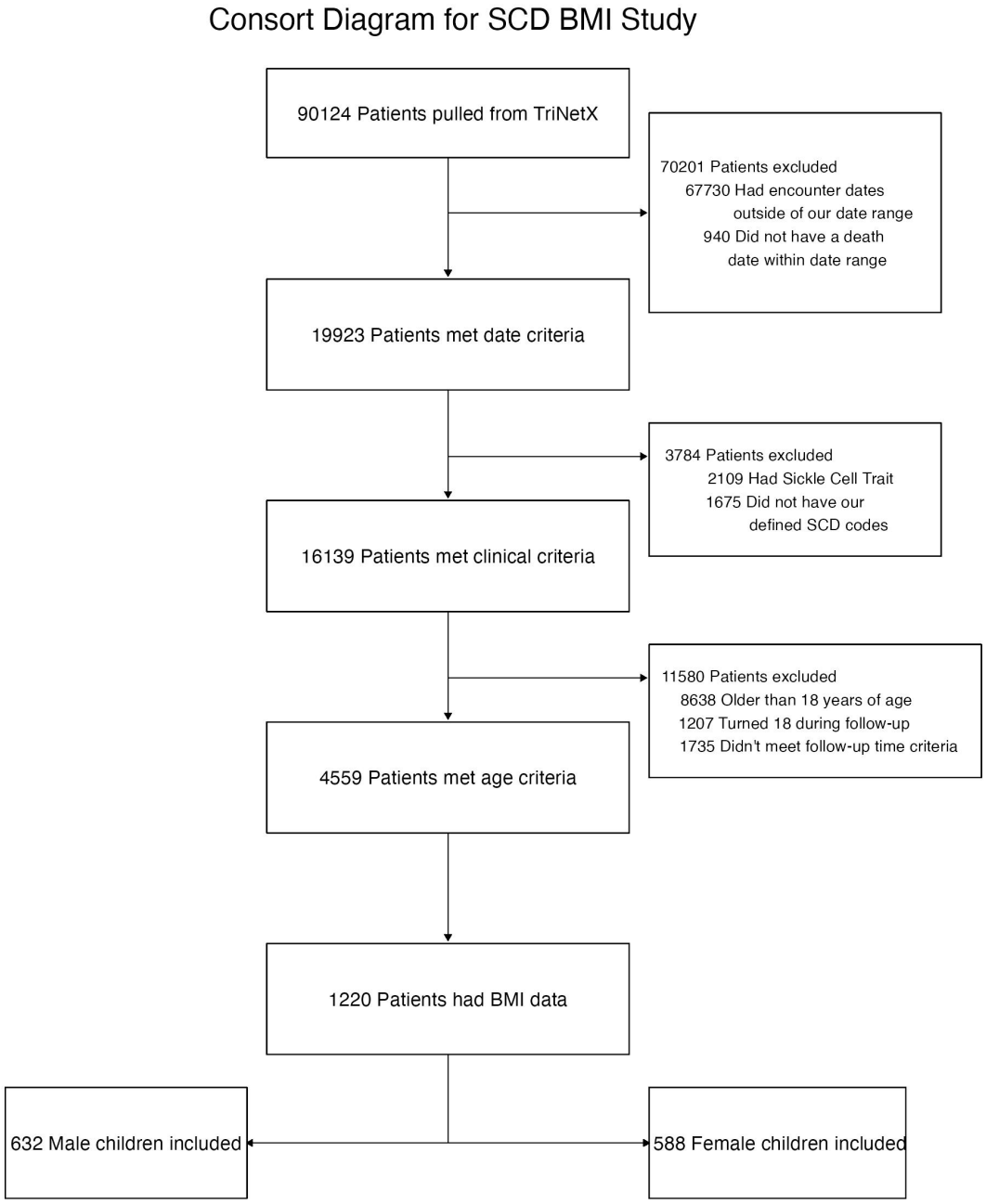

Figure 2.1 – Consort Diagram for Inclusion/Exclusion

2.4 Data Manipulation

The analytical procedures employed in this paper were conducted using the statistical software SAS (v9.4, Analytical Products 15.2) on the MCW computing cluster. This platform facilitated the comprehensive examination of the data, ensuring robust and accurate findings. Data manipulation, data management, and all statistical results were produced using the available procedures published through the SAS Institute.

2.4.1 Identifying BMI Data & Adjusting for Outliers

The identification of BMI data was achieved by querying the *Vitals Signs Table* through two primary methods. Firstly, it was directly identified utilizing the LOINC code '39156-5'. Alternatively, it was calculated based on height and weight measurements. The calculation employed the relevant LOINC codes, '8302-2' for height and '3141-9' for weight, as referenced in equation 1.1. This data was readily available in the vitals data set provided by TriNetX. Individuals were then assigned a BMI corresponding to their age.

BMI-for-age data is susceptible to misreporting for a number of reasons, with one significant factor being measurement errors. These errors often arise from inaccuracies in measuring height and weight, which are essential for calculating BMI. Such inaccuracies can be due to improper use of measuring equipment or natural fluctuations in a child's weight. In addition to measurement errors, self-reporting issues contribute significantly to data inaccuracies. This is particularly relevant when BMI data is based on self-reported height and weight, which is known to be less reliable. Individuals, especially adolescents, might report incorrect figures, either intentionally or unintentionally, influenced by their perceptions of body image or a simple misunderstanding of their actual measurements (Daniels, 2009).

To mitigate the issue of outliers/misreporting and enhance the accuracy of our analysis, we adopted a strategy of using the median BMI value for each individual for each year they reported data. This approach was chosen as it helps to reduce the impact

of extreme values or inaccuracies that might skew the results. By focusing on the median BMI per year, we aimed to obtain a more representative and reliable measure of each individual's BMI, considering the potential variations within the data.

Despite the implementation of median-based outlier mitigation techniques, our analysis revealed the persistence of some anomalous BMI values that indicated potential measurement errors. To address this, a meticulous manual review process was employed, focusing specifically on the most extreme BMI values in our dataset.

The review process entailed a detailed examination of the highest and lowest 50 BMI measurements recorded. This was achieved by constructing a comparative analysis table. The table juxtaposed each individual's BMI measurement against their BMI values recorded in the preceding and succeeding years. This comparison was critical in identifying any abrupt or unreasonable changes in BMI that could not be attributed to natural physiological developments or expected variations.

For each of these extreme cases, the change in BMI was scrutinized. If a BMI measurement exhibited a drastic change from one year to the next, or if the value appeared implausible for the individual's age, it was flagged for potential exclusion. Such drastic changes were indicative of potential data entry errors or measurement inaccuracies that could significantly distort our analysis.

Upon identification of these outliers, a decision was made to remove these specific observations from the dataset. This removal was deemed necessary to uphold the integrity and accuracy of our analysis. By eliminating these improbable data points, we aimed to ensure that our BMI data more accurately reflected the population's health status and trends, free from the distortion of measurement errors.

Obs.	**patient_id**	**age**	**prev_BMI**	**BMI**	**next_BMI**
81	7937e...	8	.	44.850	15.120
92	c4c69...	9	.	32.000	17.000

Table 2.1 – Observations Removed due to Extreme BMI Values

An intriguing aspect of our analysis was the identification of numerous extreme BMI values segmented by age. Initially, these outliers appeared to be potential data er-

rors, warranting removal from the dataset. However, we observed that many of these extreme BMI values were consistently supported by surrounding BMI measurements from the same individuals in adjacent years. This pattern of consistency suggested that these were not mere anomalies or errors, but rather legitimate instances of extreme BMI values.

2.4.2 Adjusting for Age Uncertainty

It should be noted that due to patient privacy concerns, the TriNetX database provides only the 'year of birth' rather than the precise 'date of birth'. This limitation implies that we are unable to ascertain the exact ages of individuals in our cohort. Understanding the limits of uncertainty, consider the pool of patients where the difference in their 'year of birth' and the year of their BMI measurement is 3. It's possible that a patient could have a recorded BMI measurement on the day before turning 4 years old and be pooled with individuals who turned 3 years old a day prior. This suggests there is a full year of uncertainty in our age variable. The following table demonstrates a scenario posed by our uncertainty:

Patient #	Birth Year	BMI Date	Possible Birth Date	Age by Year	Age By Birth Date
1	2014	01/02/2017	01/01/2014	3	2
2	2014	12/30/2017	12/31/2014	3	3
3	2014	01/02/2018	01/01/2014	4	3
4	2014	12/30/2018	12/31/2014	4	4

Table 2.2 – Uncertainty Example

To understand how this uncertainty effects the reliability of our outcome, we can introduce this error into the data provided by the CDC for healthy children and investigate ways to adjust. In mirroring this uncertainty to the CDC dataset, we averaged the BMI values at each age with the surrounding ages. This method effectively

broadens the age ranges in the CDC data, rendering the two datasets comparable in terms of age-related uncertainty, and ensuring a more accurate cross-analysis.

We illustrate the effect of this uncertainty by fitting a quantile regression with plots generated for both the unadjusted and adjusted CDC datasets. The disparity in position between these regression lines was then calculated to identify the ages most susceptible to error due to the simulated uncertainty. This step enabled the pinpointing of the least error-prone age segments for further analysis. This analysis suggests a focus on the age range of 8-17 for both males and females to ensure more reliable results. The graphical representations of these findings are displayed:

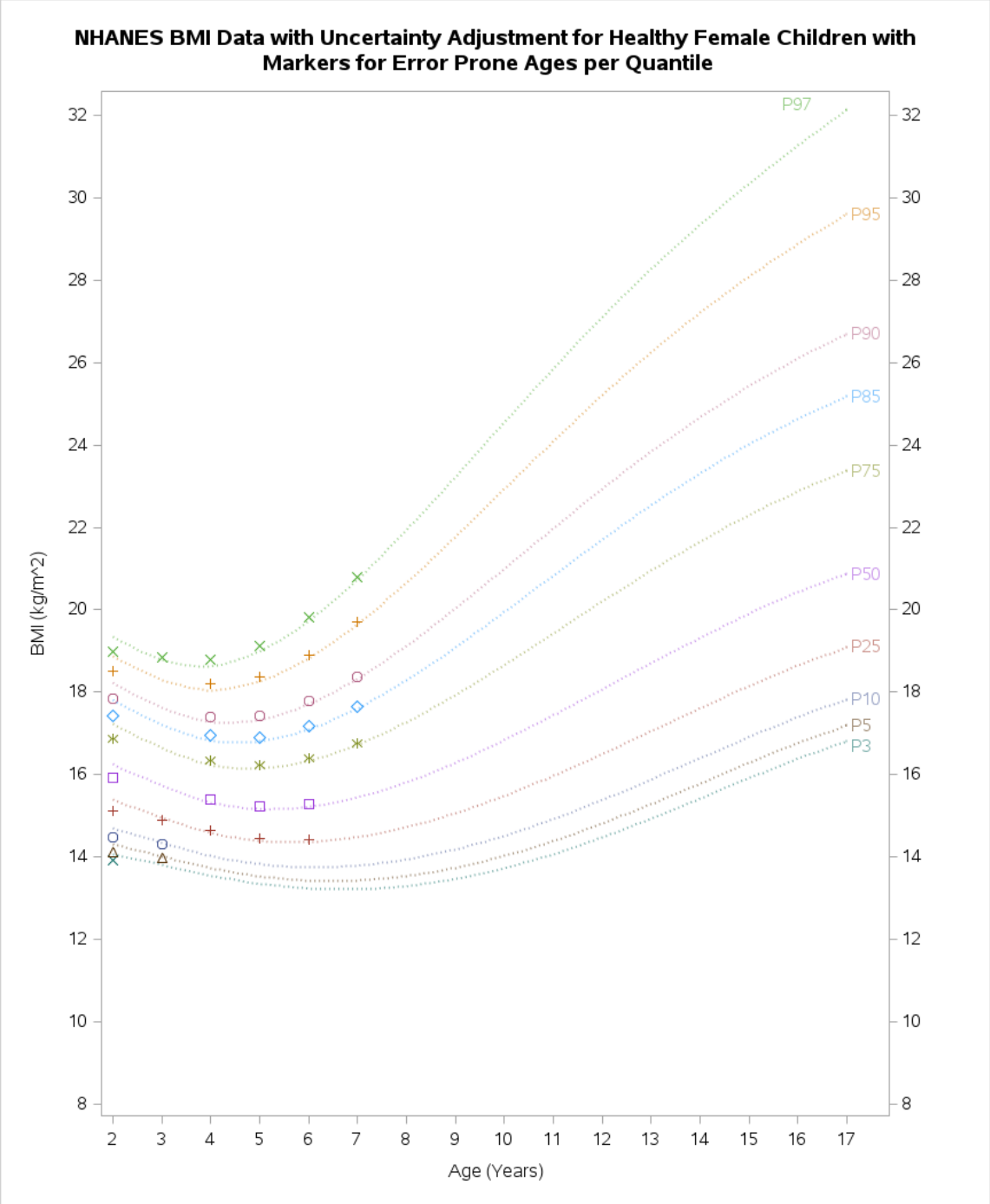

Figure 2.2 – Visualization of Error Prone Ages - Female Children

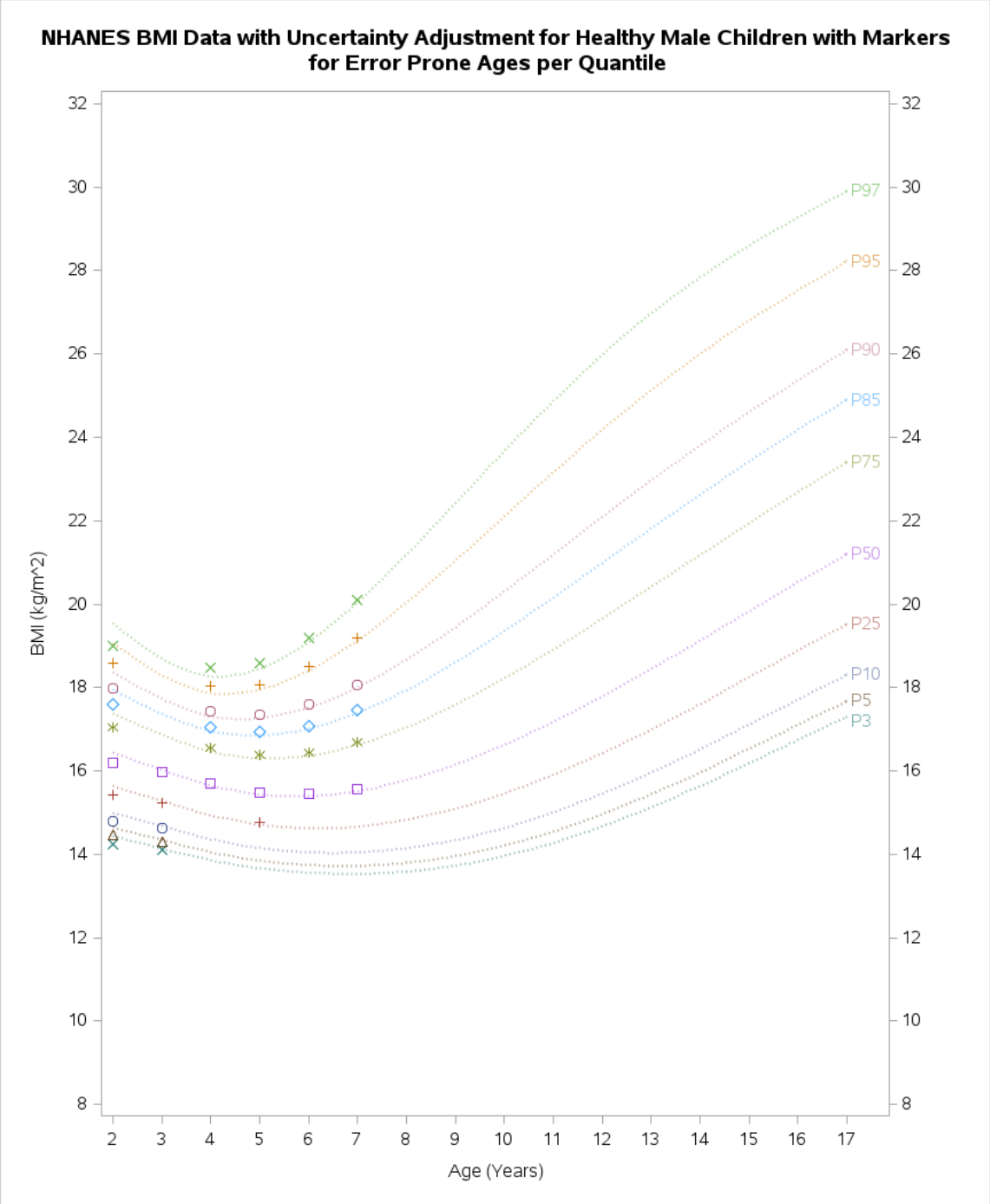

Figure 2.3 – Visualization of Error Prone Ages - Male Children

Chapter 3

Results

3.1 Cohort Demographics

Patients who met the criteria described in 2.3 were included in our cohort, and segmented by sex. A summary of demographic characteristics of the cohort is shown where the data, represented in 3.1, includes key variables such as race, ethnicity, regional location, and the distribution of the year of birth. Race is a field provided in the demographics table. For the context of our study, individuals not identified as White, Black, or Unknown were categorized as "Other" in the race variable. Regional Location categorized patients by their location in the United States. Year of Birth records the patients year of birth, and is our most specific understanding of the individual's age.

We find that our cohort yields a significant majority of patients identified as Black or African American: 79% of females and 78% of males. Other racial categories such as White and "Other" constituted smaller percentages, with White individuals making up 8.5% of females and 9.7% of males. For regional location of patients, we find that the South had the highest representation with 69% of females and 68% of males, followed by the Midwest. The West and Northeast regions had comparatively lower representation. Given the context provided in 1.3, which highlights that SCD predominantly affects African Americans, the racial demographics of this cohort align

with expectations. Further, the high representation from the South in the regional distribution is consistent with the larger African American population in that region.

Summary Statistics of SCD BMI Cohort		
Characteristic	**F**, N = 588[1]	**M**, N = 632[1]
Race		
Black or African American	462 (79%)	493 (78%)
Other	29 (4.9%)	32 (5.1%)
White	50 (8.5%)	61 (9.7%)
Unknown	47 (8.0%)	46 (7.3%)
Year of Birth		
1998-2001	5 (0.9%)	2 (0.3%)
2002-2005	108 (18%)	102 (16%)
2006-2009	94 (16%)	76 (12%)
2010-2013	97 (16%)	100 (16%)
2014-2018	284 (48%)	352 (56%)
Ethnicity		
Hispanic or Latino	34 (5.8%)	35 (5.5%)
Not Hispanic or Latino	515 (88%)	563 (89%)
Unknown	39 (6.6%)	34 (5.4%)
Regional Location		
Midwest	106 (18%)	126 (20%)
Northeast	23 (3.9%)	29 (4.6%)
South	408 (69%)	429 (68%)
West	42 (7.1%)	43 (6.8%)
Unknown	9 (1.5%)	5 (0.8%)

[1] n (%)

Table 3.1 – Cohort Demographics

3.2 Growth Chart Results

BMI-for-age growth charts were created using quantile regression for the respective Male and Female SCD cohorts. These charts delineate specific percentiles, ranging from the 3rd to the 97th, to provide a comprehensive overview of the BMI distribution across ages for children with SCD. The curves plotted on these charts are juxtaposed with those of SCD children and healthy children, offering a comparative analysis. Individual plots were also generated to show statistical significance based on where the error bars of our SCD quantile regression overlap the healthy CDC counterparts. Figures were generated in using `‘PROC QUANTREG‘` in SAS as well as `‘quantreg‘` in R for validation. Both software options yielded similar results; the figures shown in this section were produced using SAS.

The quantile regression was fit for ages 2 through 17, which encapsulates all the data available to us for our defined SCD children cohort. However, in 2.4.2, we observed that for children younger than 8 years, the 'year of birth' data caused significant discrepancies in age determination, potentially leading to skewed or inaccurate BMI assessments. By contrast, the impact of this uncertainty was markedly less in the 8-17 year age group, allowing us to provide more accurate and meaningful insights into the BMI trends within this cohort. Thus, while our quantile regression model was initially applied across a broader age range, our focus on the 8-17 year bracket in the final analysis was a deliberate choice to enhance the precision and validity of our study's findings, particularly in comparing the growth patterns of children with SCD to those of their healthy counterparts.

To facilitate the comparative analysis between our SCD cohort and the general pediatric population, error bars representing a 95% confidence interval were integrated into our quantile regression results. These error bars are crucial in quantifying the uncertainty or variability in our dataset, offering a visual gauge of the precision of our calculated BMI-for-age values. The use of a 95% confidence interval is particularly significant, as it implies that if the study were to be repeated under the same conditions, we would expect the true value to fall within this interval 95% of the time.

This addition of error bars was instrumental in our comparative analysis, enabling a direct and quantitatively robust comparison against the CDC's established growth chart lines.

Statistical significance in our study was assessed based on the overlap, or lack thereof, of these error bars with the CDC's lines. When the error bars of our quantile regression results did not intersect with the CDC's lines, it indicated a statistically significant deviation from the CDC's benchmarks. This non-overlap signifies that our observed differences in BMI-for-age are not due to random chance but are statistically significant at the 95% confidence level. On the other hand, if there was an overlap, it suggested that our results were within the expected range of variation for a healthy pediatric population, indicating no significant divergence from the CDC's benchmarks.

Employing a 95% confidence interval through error bars for visual and statistical comparison enhances the clarity and interpretability of our findings by implementing a rigorous and widely-accepted statistical basis for assessing the significance of our results. By this method, we can effectively and reliably evaluate whether our quantile regression results for BMI-for-age growth charts significantly differed from, or were in alignment with, the established growth patterns documented by the CDC.

3.2.1 Male Growth Chart

Figure 3.1 illustrates the BMI percentiles for male children with SCD in comparison to healthy male children. The x-axis represents the age range from 8 to 17 years, where this range was selected for the purpose of minimizing the uncertainty of our 'year-of-birth' variable discussed in section 2.4.2. The y-axis shows the BMI (kg/m^2) ranging from 8 to 40, where the upper percentiles approach 48. A series of colored curves represent different BMI percentiles, with labels from P3 to P97. For each percentile, there are two curves: a solid line depicting the BMI of male children with SCD and a dotted line representing the BMI of healthy male children. As age increases, the BMI curves for both SCD and healthy children show an upward trend. Notably,

in most percentiles, the BMI for male children with SCD is consistently lower than their healthy counterparts, indicating a potential growth discrepancy between the two groups. In the 95th and 97th percentiles, we find odd behavior in the plotted trends with shapes entirely different than the other percentiles. Since we've already adjusted for outliers, we investigate whether this trend is due to a lack of sample size in section 3.2.3.

The segmented figures in A.1 reveal statistical significance in the differences between our SCD group and the healthy male group at particular ages within quantiles. We find across percentiles 3, 5, 10, 95, and 97 that ages 16 and 17 show significance in the difference between groups. We also find majority ages of significant difference in percentiles 95 and 97, with some significance shown at age 14 for the 90th percentile. For the 'middle percentiles' (25, 50, 75, and 85) our figures show no significance across the age ranges. This lack of deviation gives confidence to the legitimacy of our data, as the 50th percentiles shows strikingly similar BMI-for-age prediction to the healthy group counterpart provided by the CDC.

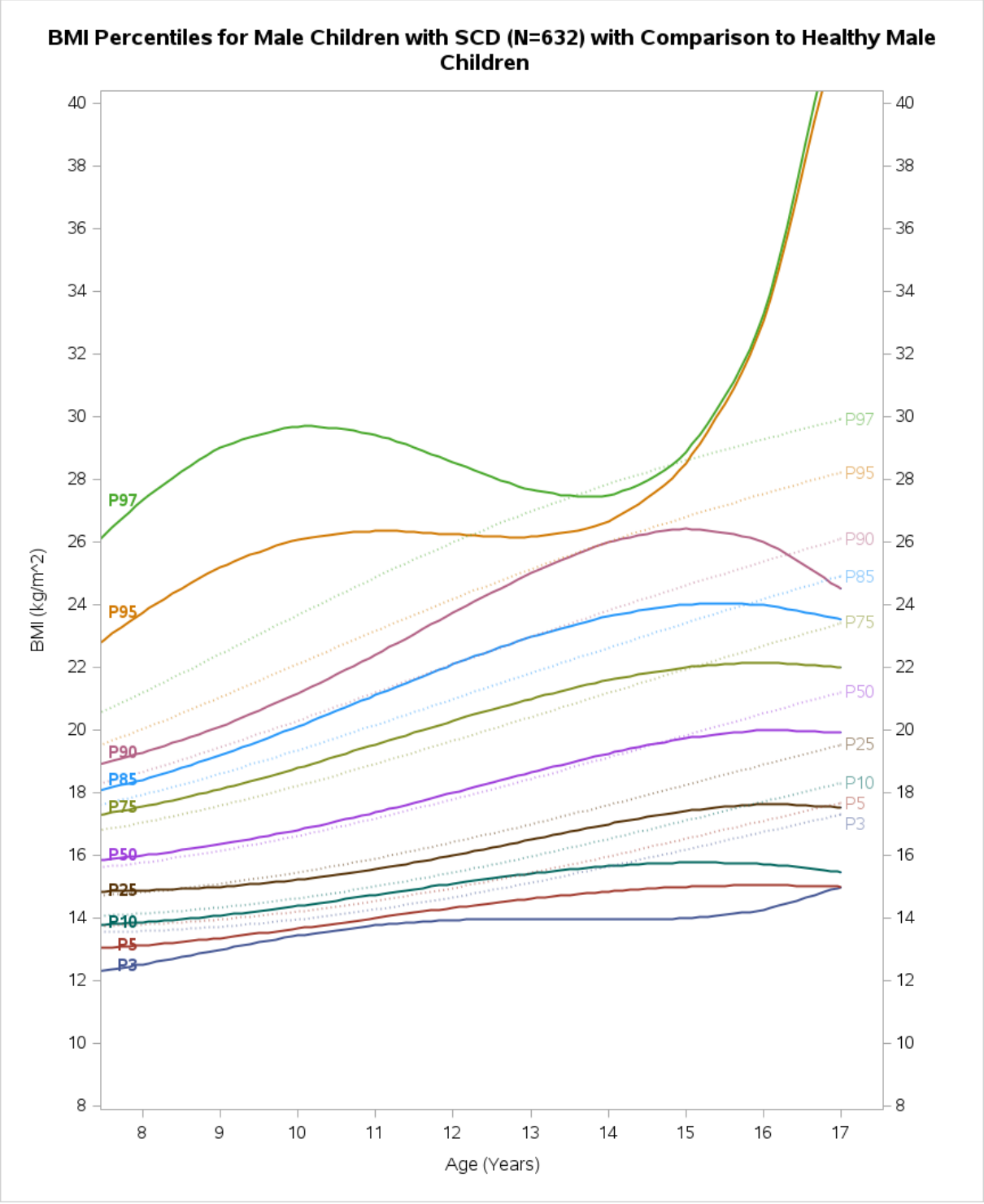

Figure 3.1 – Male Comparative BMI-for-age Growth Chart

3.2.2 Female Growth Chart

Figure 3.1 illustrates the BMI percentiles for female children with SCD in comparison to healthy female children. The x-axis represents the age range from 8 to 17 years. As with the male figure, this range was selected for the purpose of minimizing the uncertainty of our 'year-of-birth' variable discussed in section 2.4.2. The y-axis shows the BMI (kg/m^2) ranging from 8 to 40. The colored curves represent different BMI percentiles, with labels from P3 to P97. As age increases, the BMI curves for female SCD children appear to continue where the healthy female children appear to crest. This suggests that while healthy female children might experience a phase of rapid growth that later stabilizes or decreases, female children with SCD might not undergo the same stabilization, possibly due to factors associated with SCD that affect their growth patterns.

The segmented figures in A.2 reveal statistical significance in the differences between our SCD group and the healthy female group at particular ages within quantiles. As opposed to the male figures, the female group does not show significance within the early percentiles (3, 5, 10, 25), though we find significance towards the older age groups starting at percentile 50 and continuing through the rest of our percentiles. Particularly, percentiles 95 and 97 show little congruence with their healthy counterpart, as they are only within the error bounds at age 8, and deviate for the rest of the ages.

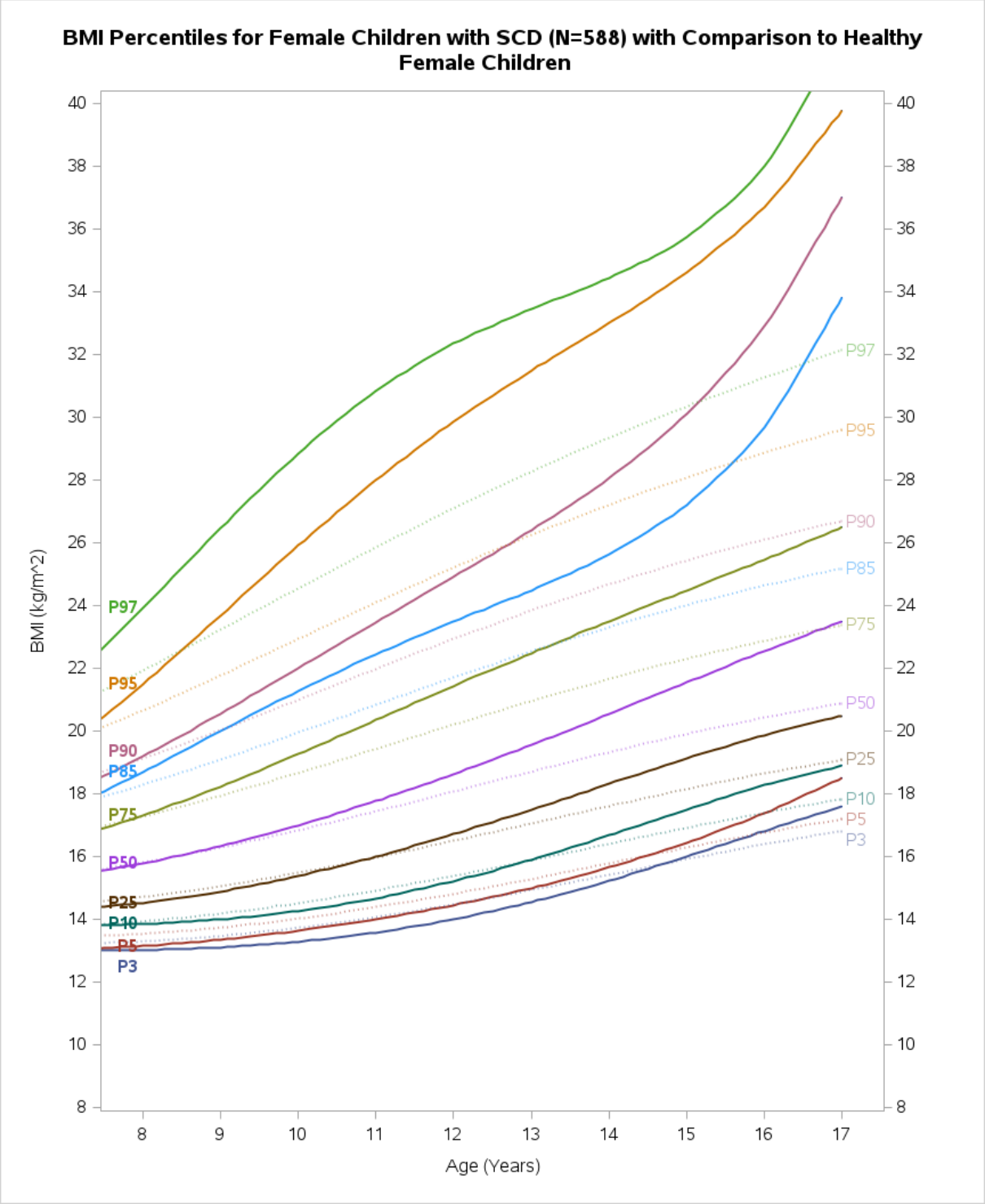

Figure 3.2 – Female Comparative BMI-for-age Growth Chart

3.2.3 Combined Sexes Growth Chart

Figure 3.3 illustrates BMI-for-age percentiles for both sexes combined with comparison to the healthy, combined CDC data. This approach was applied to discern whether the observed extremes or anomalies in our data were a consequence of a small sample size. This combined analysis was conducted across the age range of 2-17 years, acknowledging the initial similarity in BMI patterns between sexes in the early years, followed by a divergence during the teenage years, typically attributed to the onset of puberty.

The findings reveal a noteworthy trend: in the early years, the BMI data for children with SCD closely aligns with the CDC's healthy BMI data. However, as the children approach their teenage years, a divergence becomes evident. This divergence aligns with the expected physiological differences between males and females during puberty, which significantly influence BMI patterns. Though, with the data showing extremes in BMI at the upper percentiles in older children with SCD (as with the Male plot), we find that this is not merely a result of limited sample size (small N). This suggests that the variations observed in BMI among children with SCD, particularly in their teen years, are more likely indicative of the underlying health patterns specific to SCD, rather than being attributable to statistical anomalies due to sample size constraints. Further exploration should be focused on understanding the strange jump in BMI at in the higher percentiles for the age range.

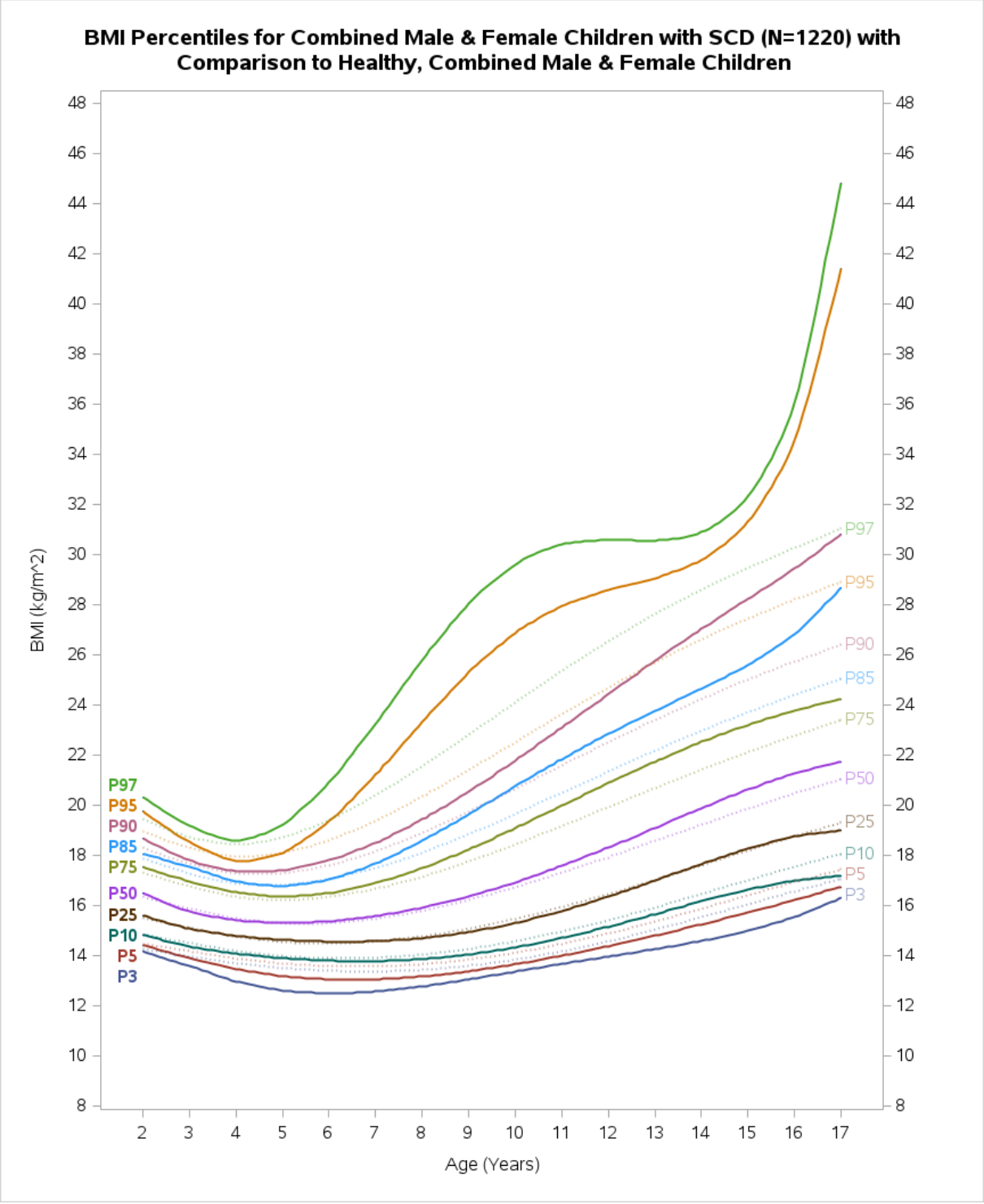

Figure 3.3 – Both Sexes Comparative BMI-for-age Growth Chart

Chapter 4

Conclusion

4.1 Summary

This capstone project presented a comprehensive analysis of BMI in children with sickle cell disease, utilizing data from the TriNetX database. Our primary objective was to examine health disparities in pediatric patients with SCD by contrasting their BMI profiles against the national BMI standards provided by the CDC.

In the initial sections of this capstone, we delved into the historical context and medical significance of SCD, alongside a detailed exposition on the concept and relevance of BMI in pediatric health. The methodological framework of this research was outlined, highlighting our inclusion and exclusion criteria that underpinned the formation of a patient cohort from the TriNetX database.

The heart of our analytical approach was the application of quantile regression. This statistical method enabled us to capture the nuances and variability in BMI distributions across different quantiles, thereby offering a more granular understanding of BMI trends within the pediatric SCD population. This approach was particularly crucial in acknowledging the non-uniform impact of SCD on body weight and composition.

The results of our study were illustrated through comprehensive charts, juxtaposing

the BMI data of children with SCD against the CDC's BMI benchmarks for children. Our findings highlighted the disparities, providing insights into how SCD can differentially impact the growth patterns and nutritional status of affected children. These disparities underscore the necessity for healthcare interventions in this vulnerable group, taking into account the unique challenges posed by SCD.

Through this research, we contribute to the emerging body of knowledge on pediatric SCD and BMI. The findings are presented to future studies which seek to utilize large clinical databases like TriNetX to explore more nuanced aspects of health and nutrition in chronic pediatric conditions. Furthermore, findings contribute to the available tools for pediatric healthcare providers, offering evidence-based insights to inform better clinical practices geared towards improving the health outcomes of children living with SCD.

4.2 Challenges and Limitations

4.2.1 TriNetX Data Reliability

One of the significant challenges we encountered in this research was the inherent limitations of the TriNetX database. TriNetX, as a source, presents raw data that often requires extensive validation to ensure reliability. We observed that the raw data from TriNetX was somewhat unreliable, necessitating the implementation of rigorous validation techniques. These validation processes, while crucial for ensuring data accuracy, led to a notable reduction in our sample size. Despite these efforts, there remained an undercurrent of uncertainty regarding the ultimate validity of our data. This concern was particularly pronounced in our exploration of the rate of cardiopulmonary and renal complications within our pediatric SCD cohort. For instance, TriNetX reported seemingly unreasonable rates of chronic illnesses such as chronic kidney disease, cardiomyopathy, or heart failure. These reports raised doubts about the data's reliability, leading us to question the authenticity of such high rates of chronic conditions in this population.

The way TriNetX collects data may contribute to these challenges. The platform aggregates patient data from a network of healthcare organizations, including electronic health records and insurance claims. This method of data collection, while expansive, can introduce inconsistencies and errors. The data is not only subject to the variability in how different institutions record and report health information but also to the inherent biases in the kinds of patients who are more frequently represented in healthcare systems or have higher insurance interactions. Consequently, the data may not accurately represent the broader patient population, especially for less common conditions like SCD.

This is compounded with how TriNetX lacks specific birth date information for patients (discussed in 2.4.2). The database only provided the 'year of birth,' rather than precise birth dates. This limitation imposes a significant constraint on accurately determining the exact ages of children at the time of their BMI measurements. These factors combined make the interpretation and generalization of findings from TriNetX data quite complex, underscoring the need for cautious and critical analysis in research utilizing this resource.

4.2.2 EMR for Research Purposes

A common issue in healthcare research is the primary purpose of EMR systems (Holmes et al., 2021). EMRs are predominantly designed for clinical management and billing purposes, rather than for research. This aspect of EMRs has important implications for the type and accuracy of data they contain, which in turn affects research relying on this data, such as ours using the TriNetX database. EMRs are structured to efficiently capture and process information crucial for patient care management and to facilitate billing procedures. As a result, they often emphasize data related to procedures and interventions that have direct billing implications. This focus means that procedures are usually well-documented and meticulously recorded, as they directly correspond to the financial aspects of healthcare provision. On the other hand, aspects like laboratory results or detailed diagnostic information, which

might not have immediate billing relevance, can sometimes be less prioritized or not as accurately recorded in EMRs. For instance, diagnostic codes in EMRs may not always reflect the most current or accurate clinical understanding of a patient's condition, as they are often entered with the primary goal of fulfilling administrative and billing requirements rather than for clinical research accuracy.

This discrepancy in the accuracy and emphasis of different types of data within EMRs presents a challenge for researchers. In the context of our study, while procedure data might be robust, there could be concerns about the completeness and precision of lab results or diagnostic information. This situation necessitates additional validation and critical assessment of the data extracted from EMRs, especially when it is used for research purposes like understanding disease patterns or treatment outcomes, where accuracy and detail are paramount.

List of References

AAFP (2008). American academy of family physicians: Screening for sickle cell disease. https://www.aafp.org/pubs/afp/issues/2008/0501/p1300.html. Accessed: November 10, 2023.

AHA (2023). BMI in children. https://www.heart.org/en/healthy-living/healthy-eating/losing-weight/bmi-in-children. Accessed: November 10, 2023.

Aidoo, M., Terlouw, D. J., Kolczak, M. S., McElroy, P. D., ter Kuile, F. O., Kariuki, S., et al. (2002). Protective effects of the sickle cell gene against malaria morbidity and mortality. *The Lancet*, 359(9314):1311–1312.

Austin, P. C., Tu, J. V., Daly, P. A., and Alter, D. A. (2005). The use of quantile regression in health care research: a case study examining gender differences in the timeliness of thrombolytic therapy. *Statistics in medicine*, 24(5):791–816.

Brown, B. J., Okoko, A. R., and Ariba, A. J. (2016). Psychosocial burden of sickle cell disease on the family, nigeria. *South African Journal of Child Health*, 10(2):104–107.

Cade, B. S. and Noon, B. R. (2003). A gentle introduction to quantile regression for ecologists. *Frontiers in Ecology and the Environment*, 1(8):412–420.

CDC (2023). About children's BMI. https://www.cdc.gov/healthyweight/assessing/bmi/childrens_bmi/about_childrens_bmi.html. Accessed: November 10, 2023.

CDC (2023). CDC Growth Charts. Accessed: 10/14/2023.

CDC (2023). Complications and treatments of sickle cell disease. https://www.cdc.gov/ncbddd/sicklecell/complications.html. Accessed: November 10, 2023.

Centers for Disease Control and Prevention (2023). Data file for the extended cdc bmi-for-age growth charts for children and adolescents. The Extended CDC BMI-for-age growth charts use a new method for calculating BMI percentiles and z-scores.

Children's Hospital of Philadelphia (2023). Sickle Cell Disease in Children. Accessed: 10/14/2023.

Cole, T. J. and Green, P. J. (1992). Smoothing reference centile curves: the LMS method and penalized likelihood. *Statistics in medicine*, 11(10):1305–1319.

Daniels, S. R. (2009). The use of bmi in the clinical setting. *Pediatrics*, 124(Supplement_1):S35–S41.

DeBaun, M. R. and Telfair, J. (2012). Sickle cell disease: A review for the pediatric dentist. *Pediatric Dentistry*, 34(2):159–167.

Fournier, J.-M. and Koske, I. (2012). The determinants of earnings inequality: evidence from quantile regressions. *OECD Journal: Economic Studies*, 2012(1).

Hall, R., Gardner, K., Rees, D., et al. (2018). High body mass index in children with sickle cell disease: a retrospective single-centre audit. *BMJ Paediatrics Open*, 2:e000302.

Holmes, J. H., Beinlich, J., Boland, M. R., Bowles, K. H., Chen, Y., Cook, T. S., Demiris, G., Draugelis, M., Fluharty, L., Gabriel, P. E., et al. (2021). Why is the electronic health record so challenging for research and clinical care? *Methods of information in medicine*, 60(01/02):032–048.

Jackson, E., Karlson, C., Herring, W., Okhomina, V., Lim, C., Morrow, A., Daggett, C., Arnold, L., and McNaull, M. (2022). Prevalence of raised body mass index in paediatric sickle cell disease. *J Paediatr Child Health*, 58(10):1829–1835. Epub 2022 Jul 13. PMID: 35822947.

Kavanagh, P. L., Fasipe, T. A., and Wun, T. (2022). Sickle cell disease: A review. *JAMA*, 328(1):57–68.

Koenker, R. and Bassett, G. (1978). Regression quantiles. *Econometrica*, 46(1):33–50. Accessed 31 Oct. 2023.

Machado, J. A. and Silva, J. (2013). Quantile regression and heteroskedasticity. *Unpublished manuscrip, Department of Economics, University of Essex, available at*.

Makani, J., Ofori-Acquah, S. F., Nnodu, O., Wonkam, A., and Ohene-Frempong, K. (2013). Sickle cell disease: New opportunities and challenges in africa. *The Scientific World Journal*.

Meinshausen, N. and Ridgeway, G. (2006). Quantile regression forests. *Journal of machine learning research*, 7(6).

Mpalampa, L., Ndugwa, C., Ddungu, H., and Idro, R. (2012). Foetal haemoglobin and disease severity in sickle cell anaemia patients in kampala, uganda. *BMC Blood Disorders*, 12:11. PMID: 22958547; PMCID: PMC3520739.

Ohio State University (2023). [https://cancer.osu.edu/for-patients-and-caregivers/learn-about-cancers-and-treatments/cancers-conditions-and-treatment/benign-blood-diseases/sickle-cell-anemia]. Accessed: 10-14-2023.

Rees, D. C., Williams, T. N., and Gladwin, M. T. (2010). Sickle-cell disease. *The Lancet*, 376(9757):2018–2031.

Stasinopoulos, D. M. and Rigby, R. A. (2008). Generalized additive models for location scale and shape (gamlss) in r. *Journal of Statistical Software*, 23:1–46.

Upadhye, D. S., Jain, D. L., Trivedi, Y. L., Nadkarni, A. H., Ghosh, K., and Colah, R. B. (2016). Neonatal screening and the clinical outcome in children with sickle cell disease in central india. *PLoS One*, 11(1):e0147081.

Walters, M. C., Patience, M., Leisenring, W., Eckman, J. R., Scott, J. P., Mentzer, W. C., Davies, S. C., Ohene-Frempong, K., Bernaudin, F., Matthews, D. C., Storb, R., and Sullivan, K. M. (1996). Bone marrow transplantation for sickle cell disease. *New England Journal of Medicine*, 335(6):369–376.

WHO (2023). Malaria. https://www.who.int/news-room/fact-sheets/detail/malaria. Accessed: November 10, 2023.

Yang, J., Meng, X., and Mahoney, M. (2013). Quantile regression for large-scale applications. In *International Conference on Machine Learning*, pages 881–887. PMLR.

Appendix A

A.1 Male Percentile Plots with Error Bars

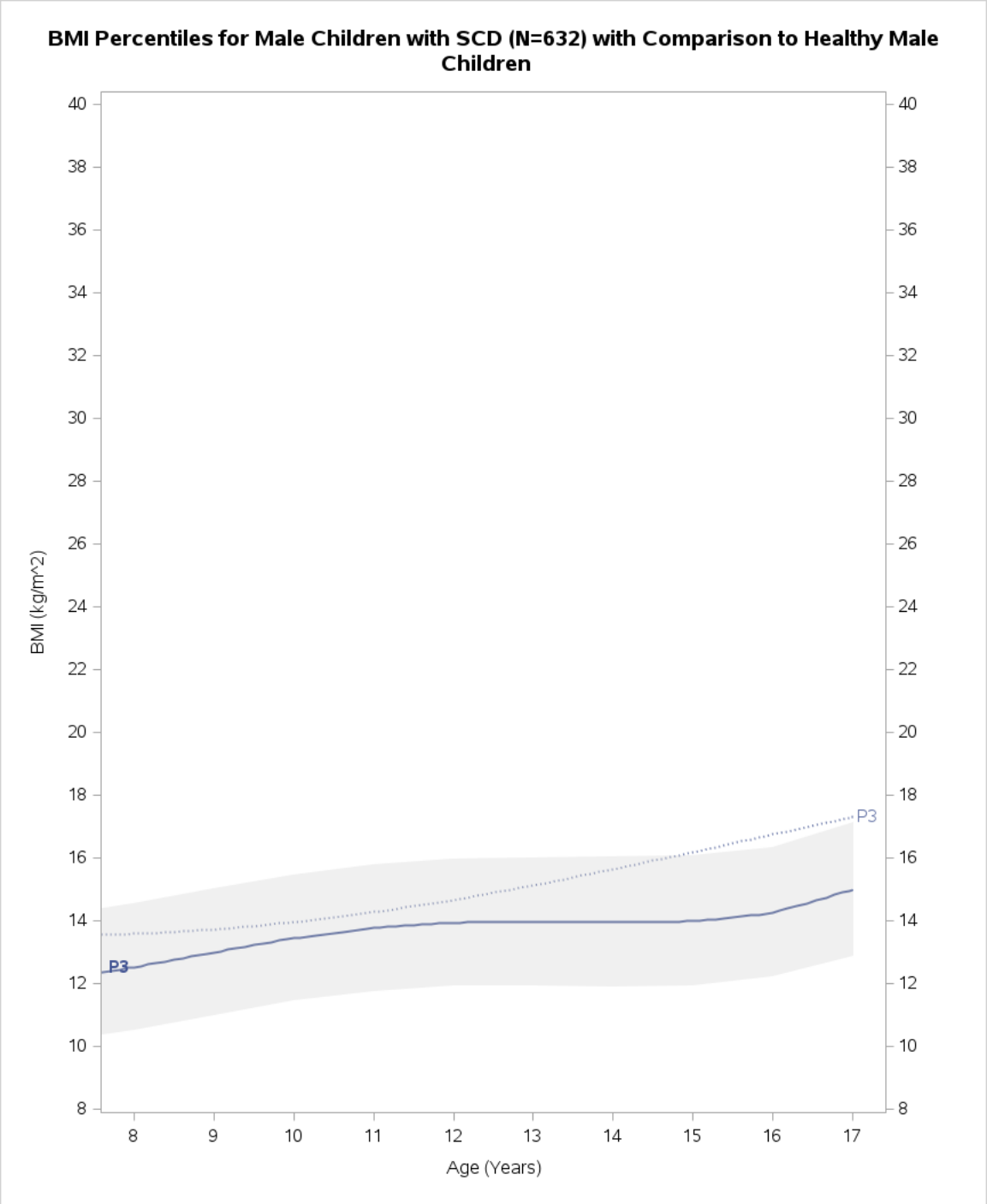

Figure A.1 – Male Growth Chart - 3rd Percentile with Error Bars

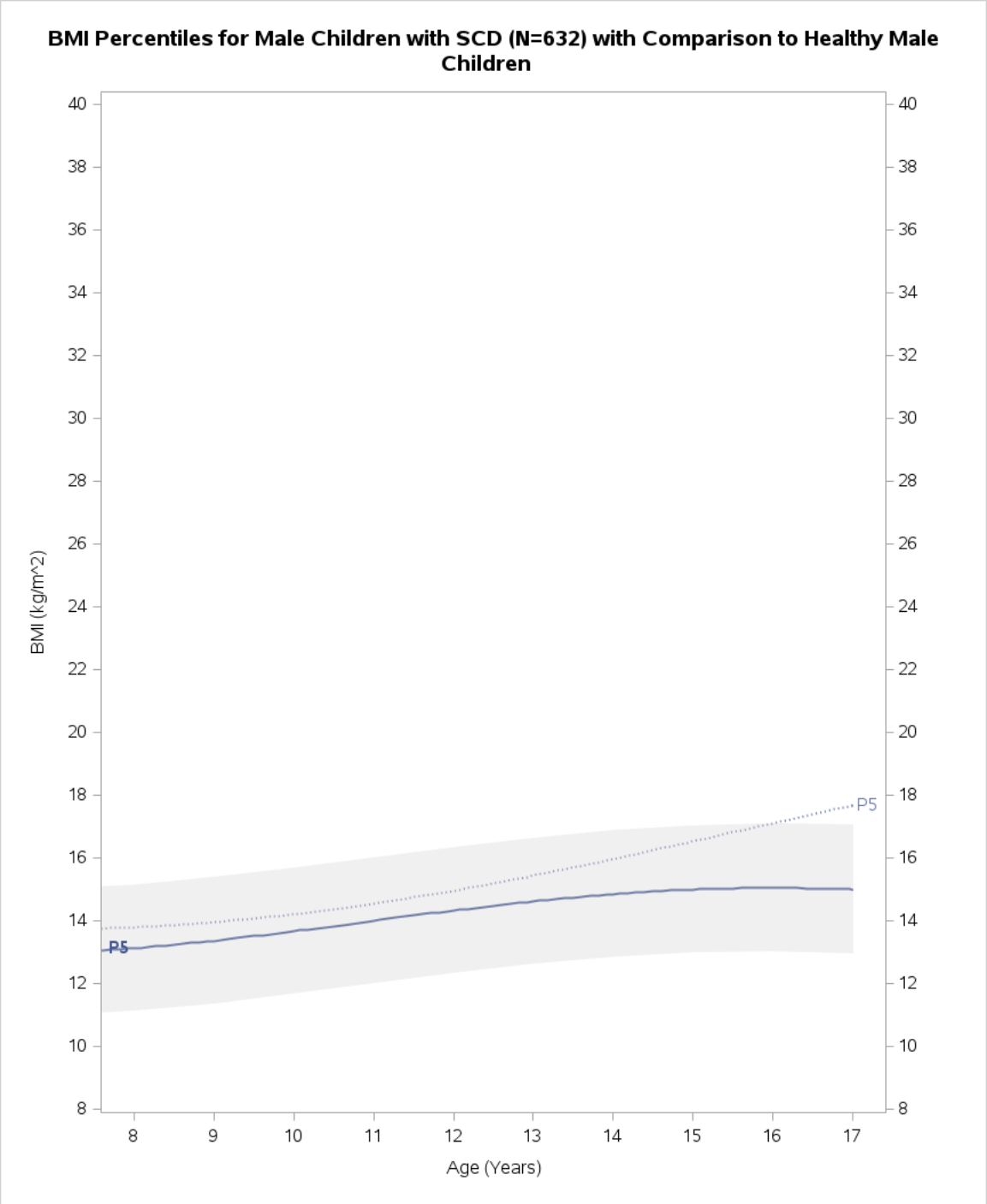

Figure A.2 – Male Growth Chart - 5th Percentile with Error Bars

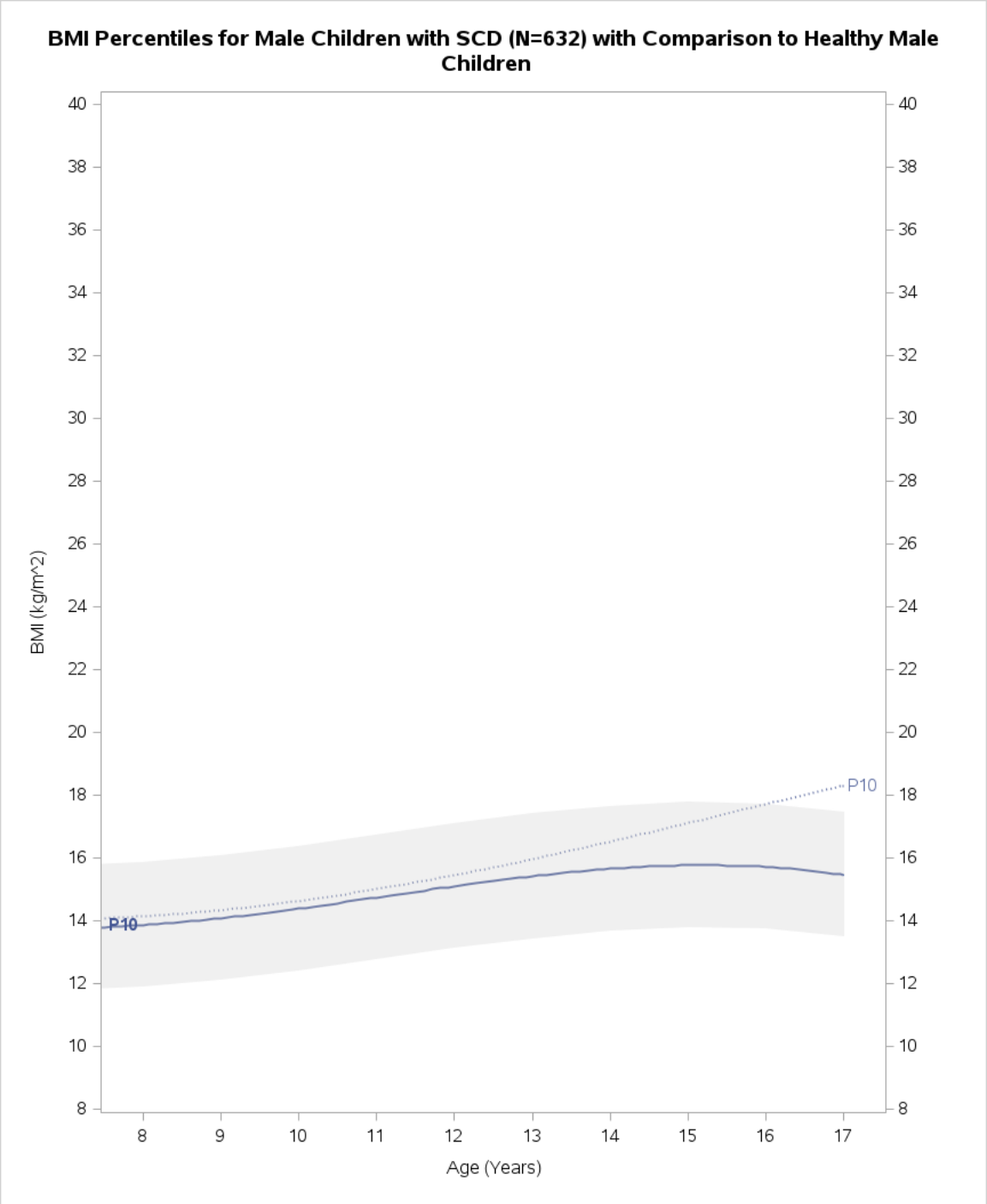

Figure A.3 – Male Growth Chart - 10th Percentile with Error Bars

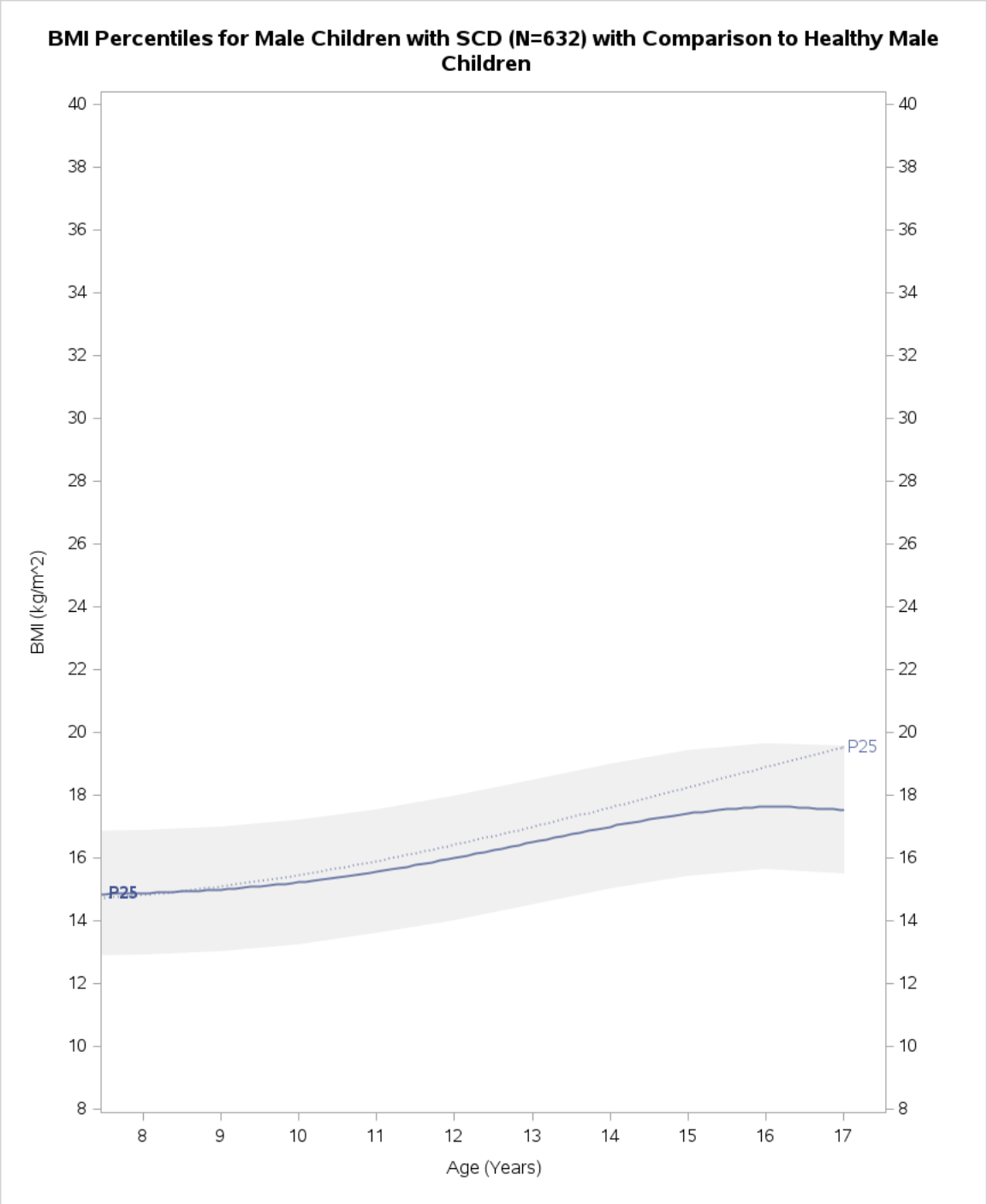

Figure A.4 – Male Growth Chart - 25th Percentile with Error Bars

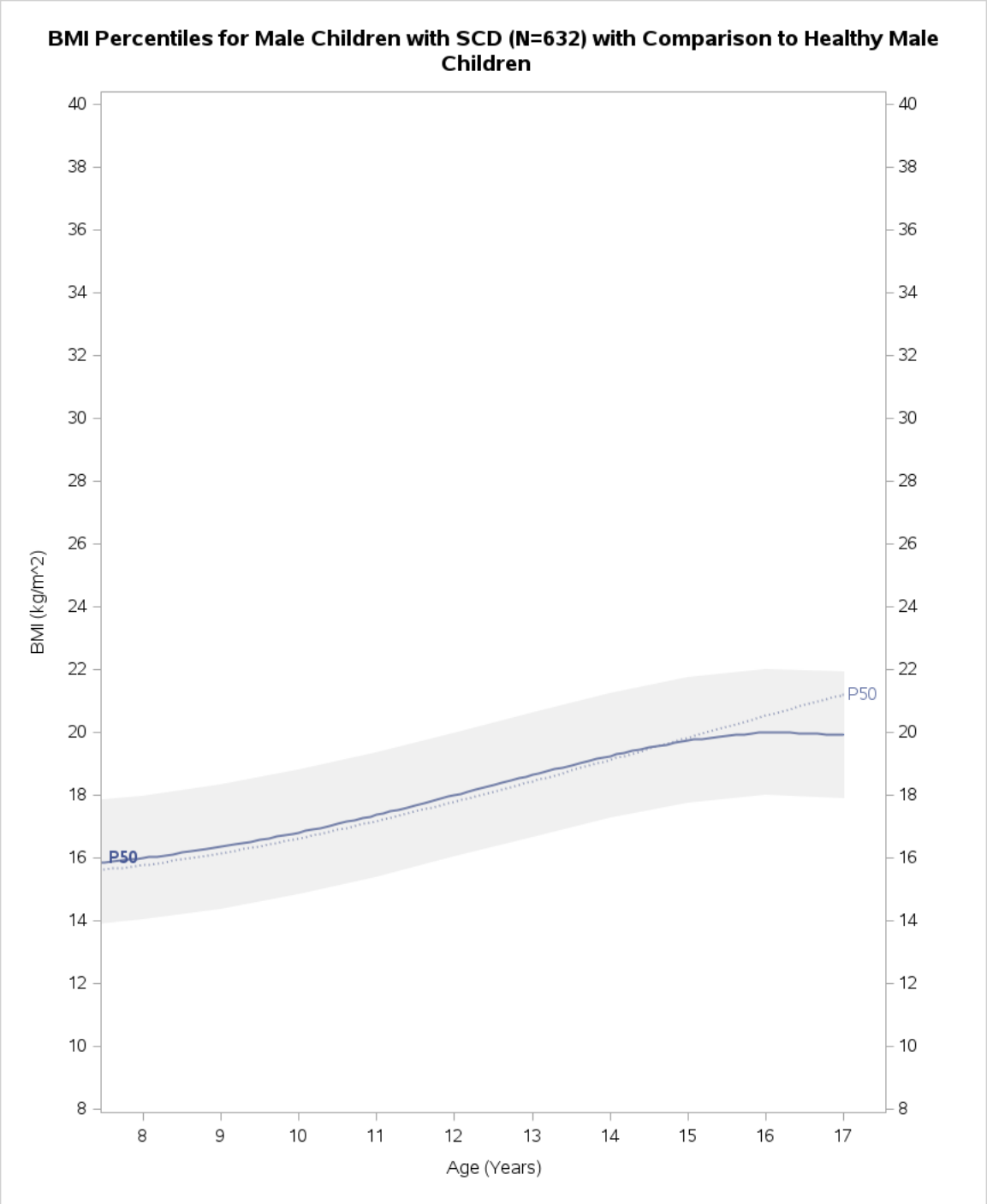

Figure A.5 – Male Growth Chart - 50th Percentile with Error Bars

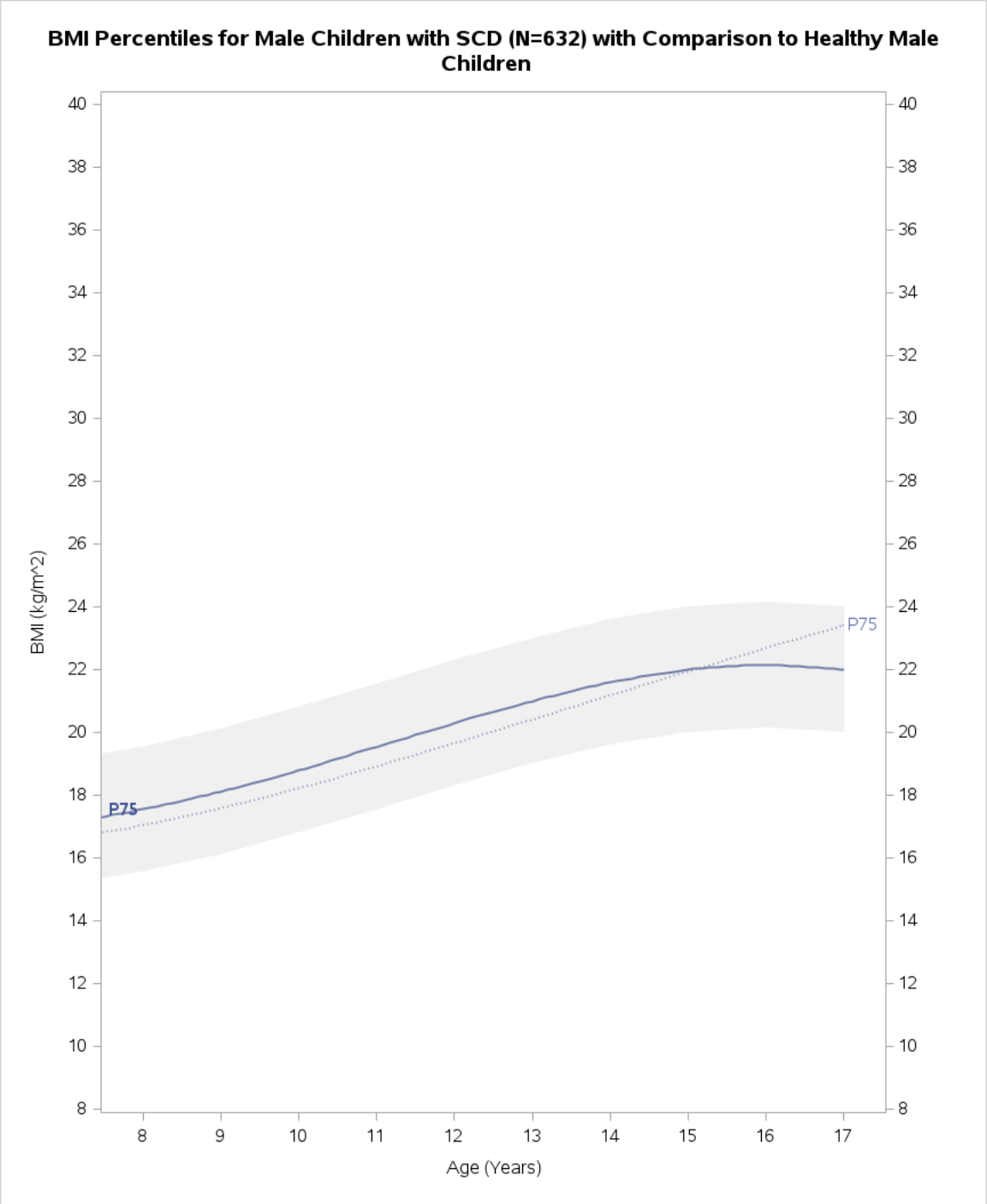

Figure A.6 – Male Growth Chart - 75th Percentile with Error Bars

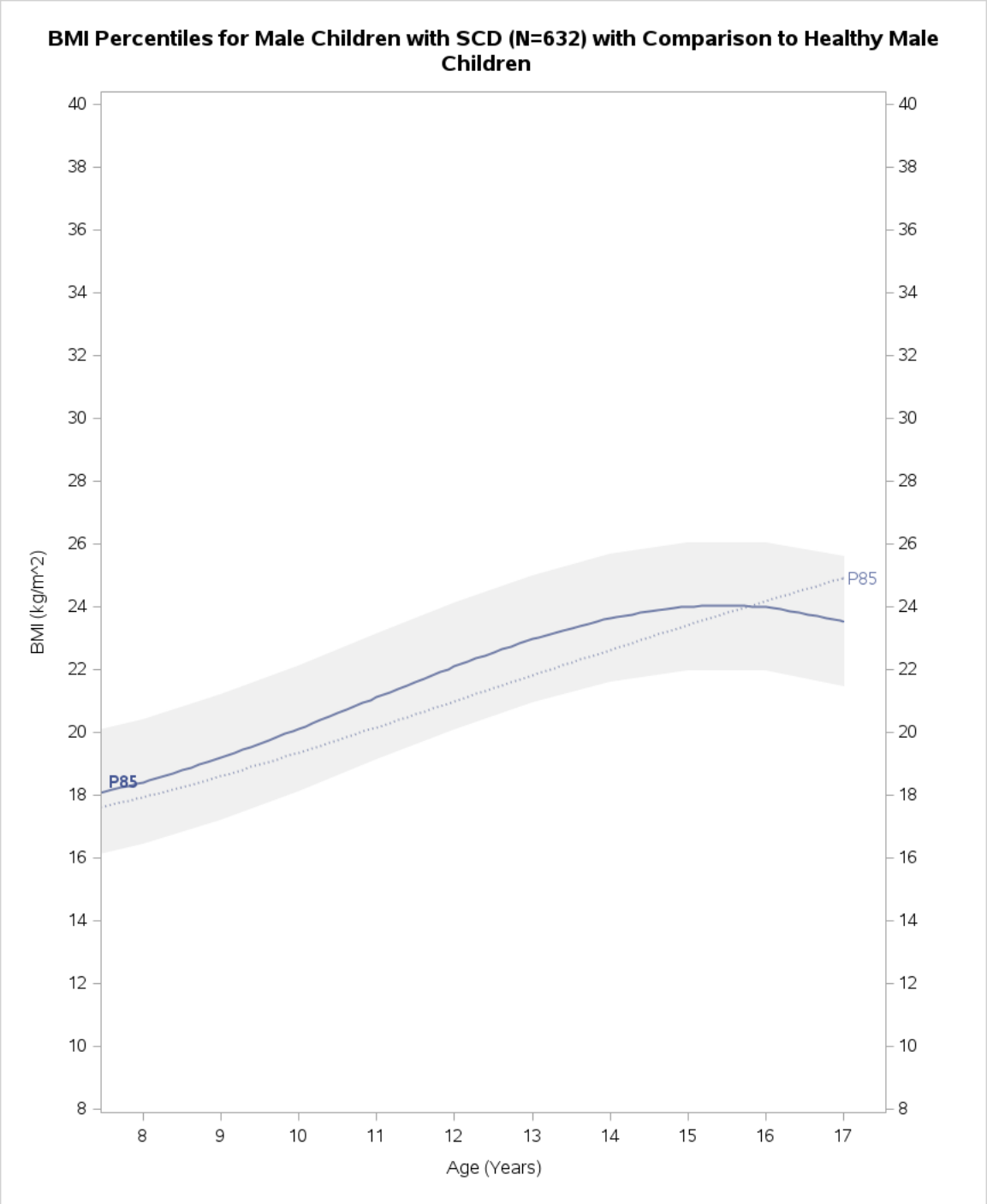

Figure A.7 – Male Growth Chart - 85th Percentile with Error Bars

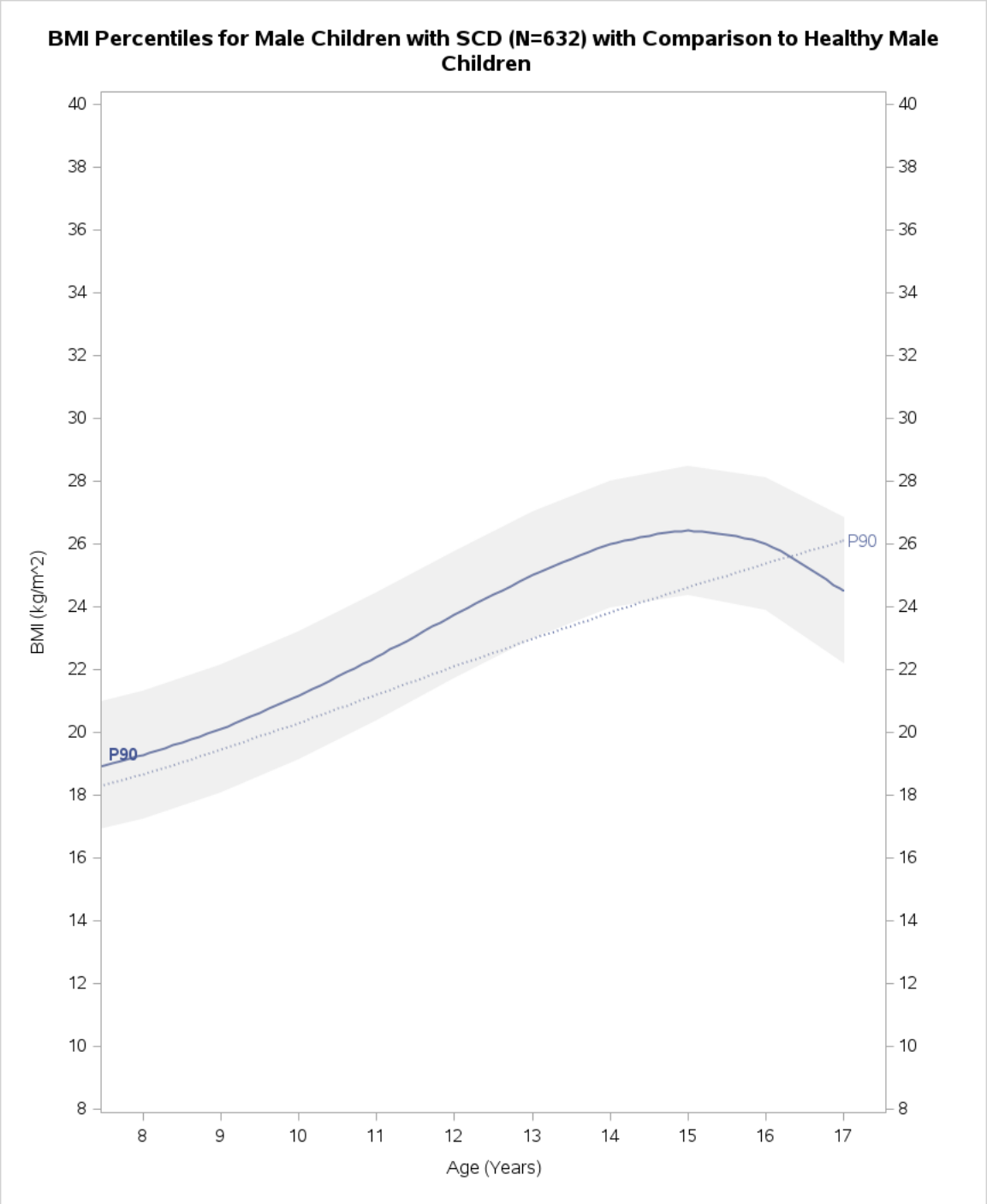

Figure A.8 – Male Growth Chart - 90th Percentile with Error Bars

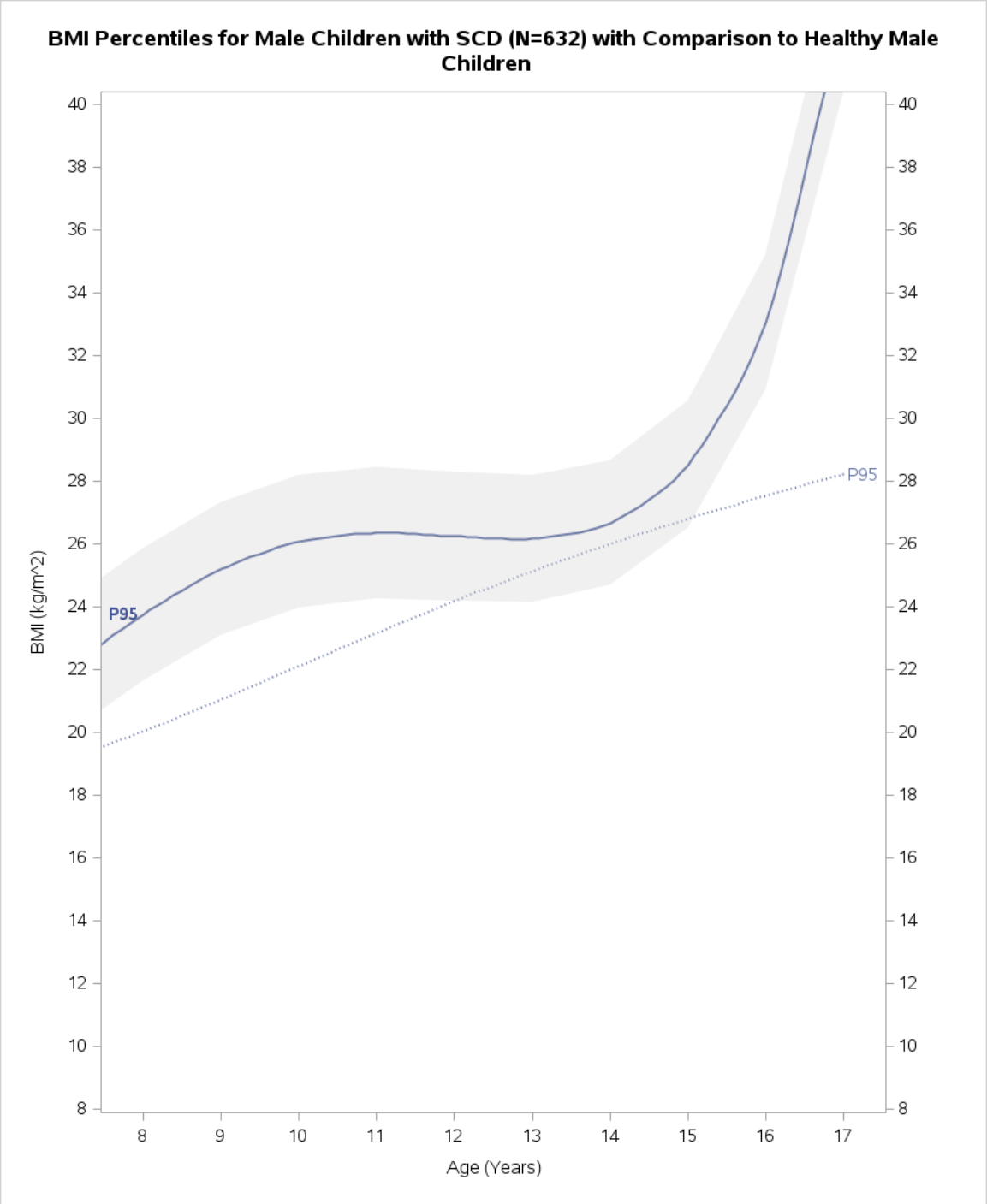

Figure A.9 – Male Growth Chart - 95th Percentile with Error Bars

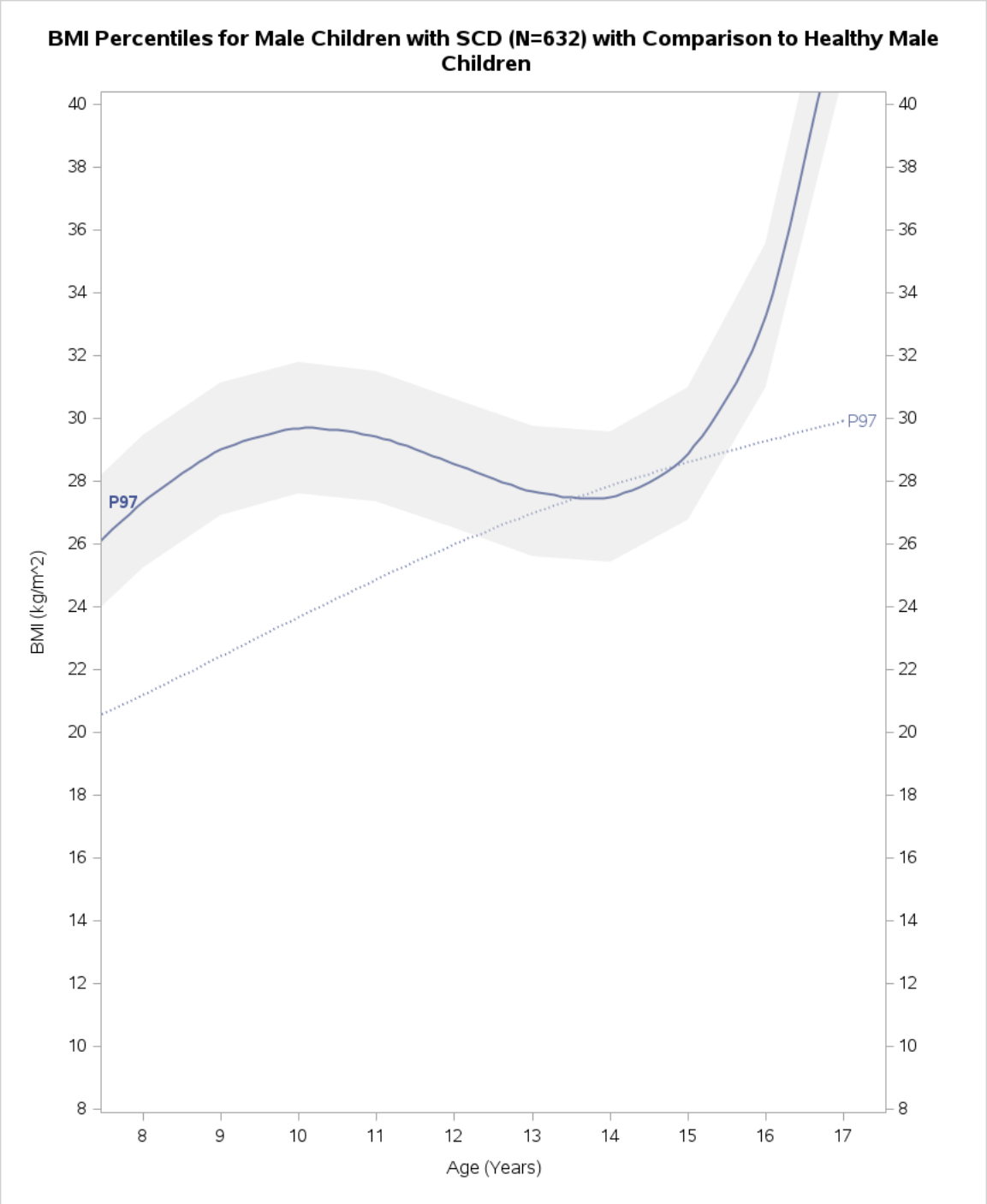

Figure A.10 – Male Growth Chart - 97th Percentile with Error Bars

A.2 Female Percentile Plots with Error Bars

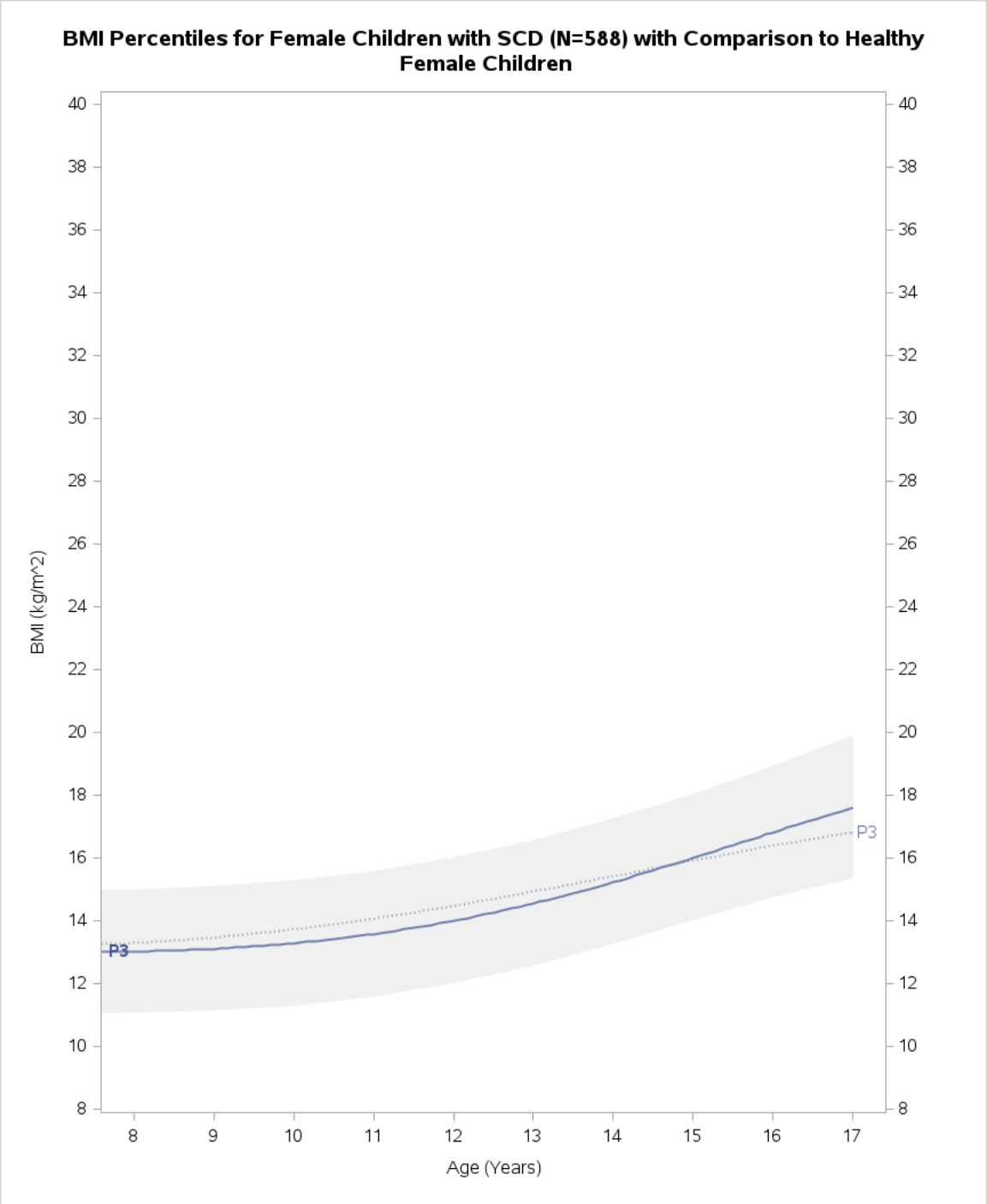

Figure A.11 – Female Growth Chart - 3rd Percentile with Error Bars

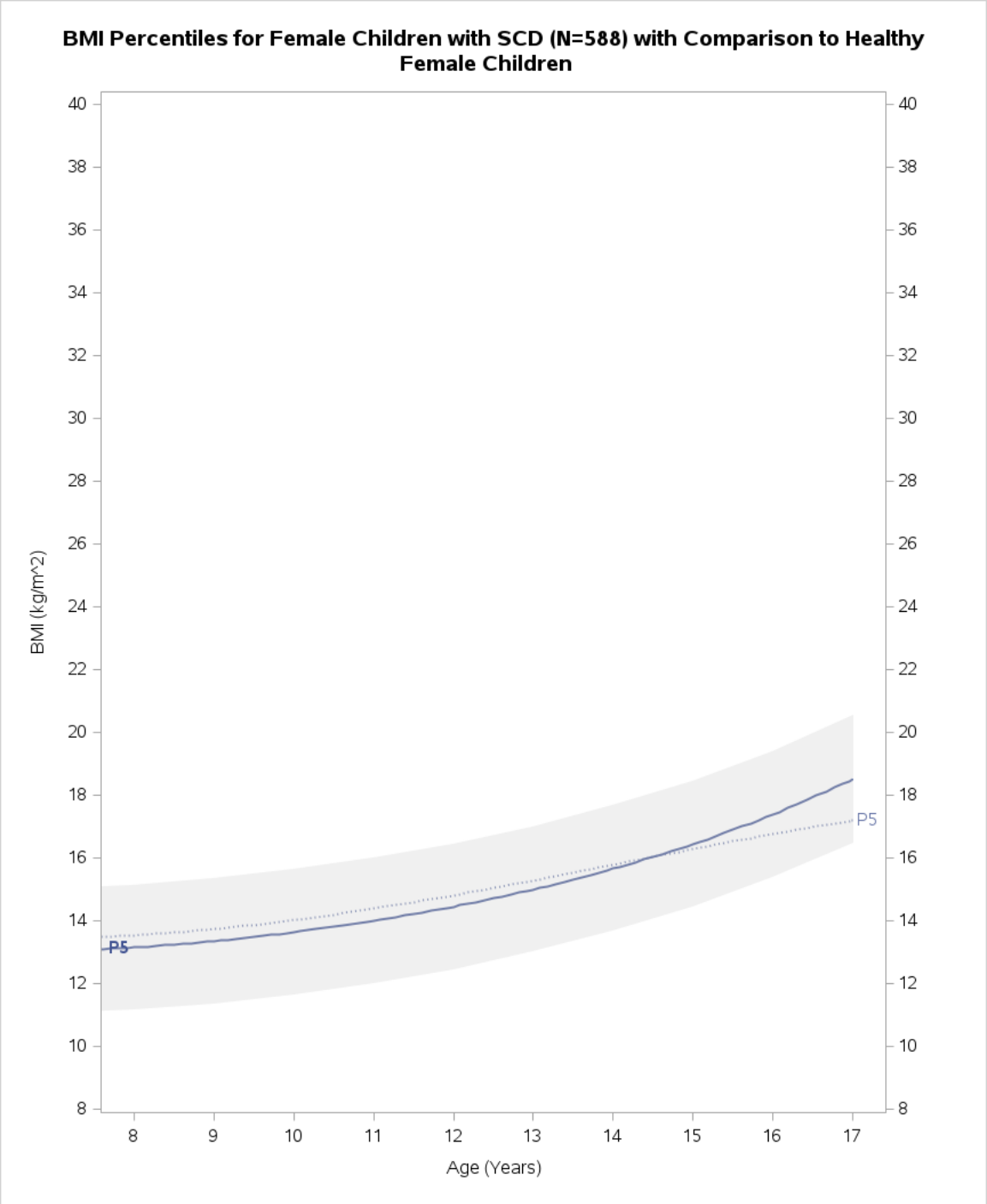

Figure A.12 – Female Growth Chart - 5th Percentile with Error Bars

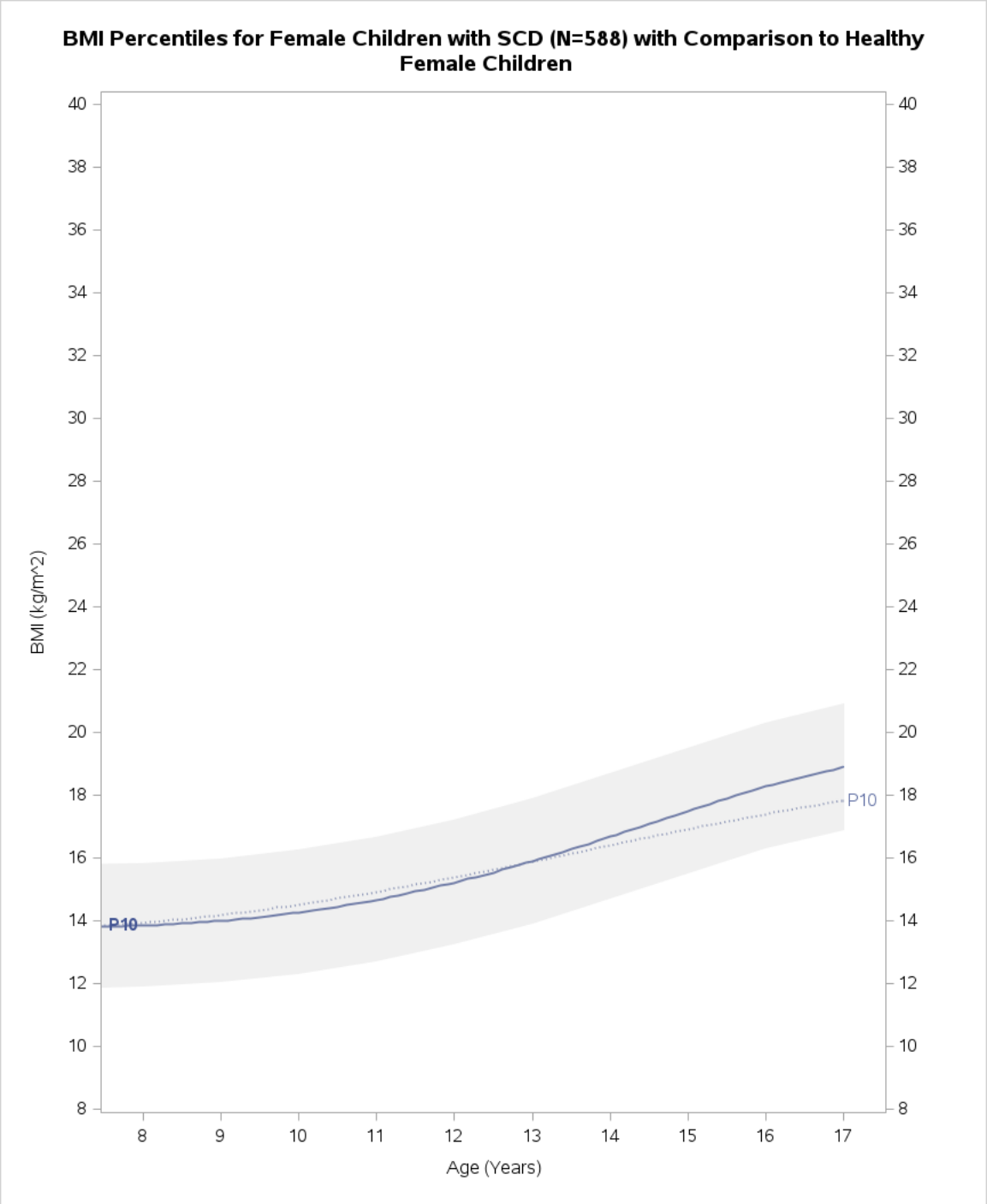

Figure A.13 – Female Growth Chart - 10th Percentile with Error Bars

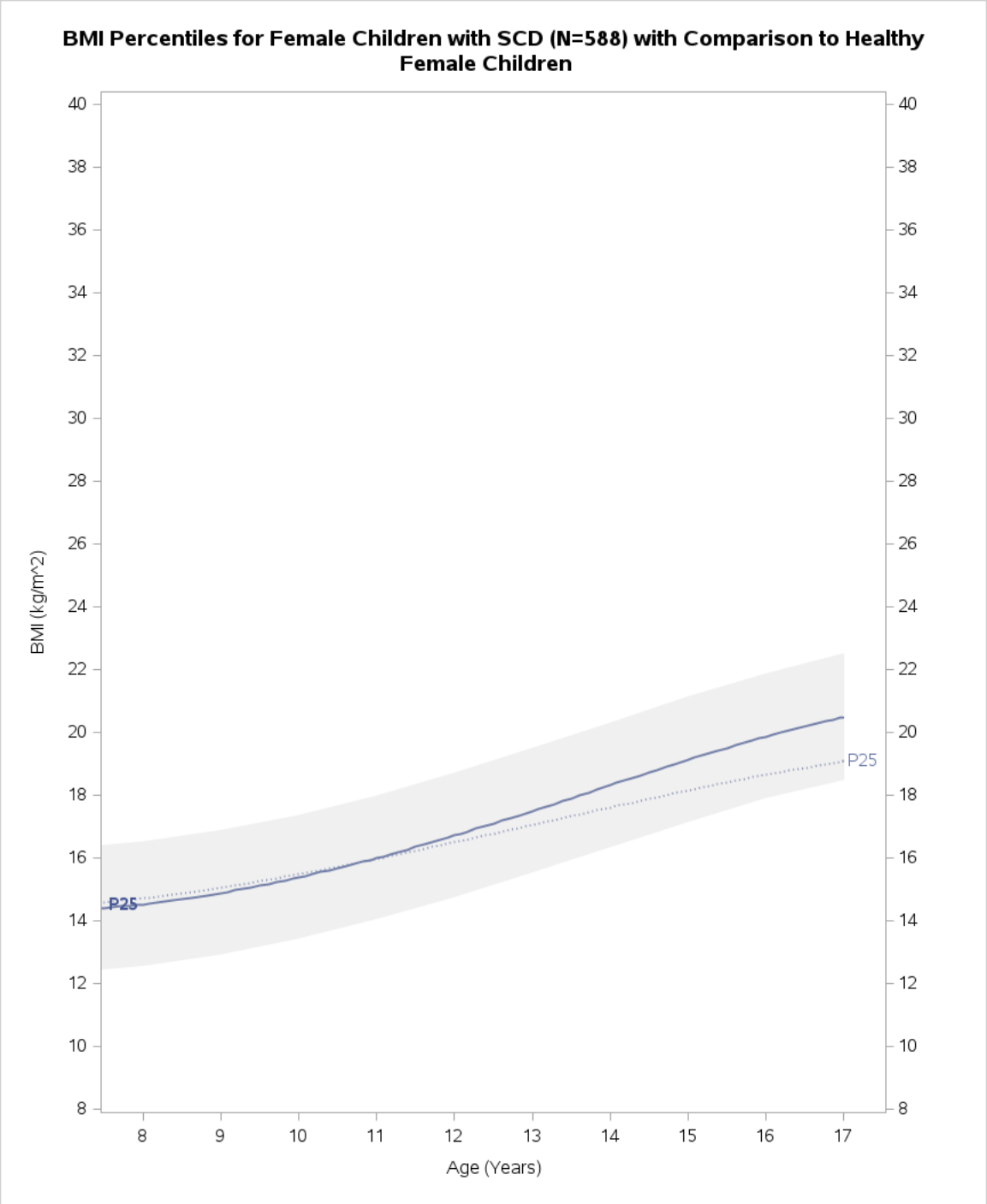

Figure A.14 – Female Growth Chart - 25th Percentile with Error Bars

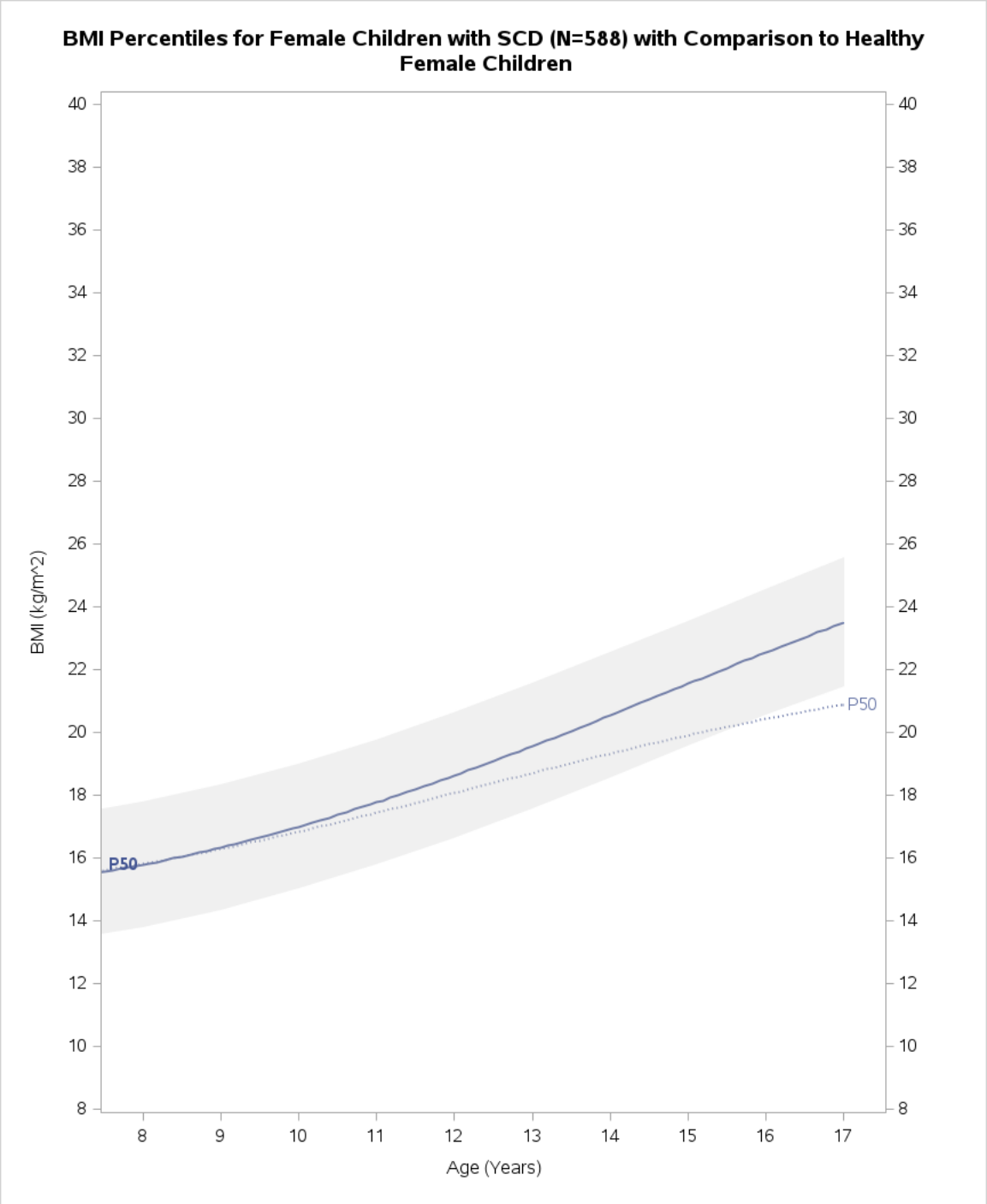

Figure A.15 – Female Growth Chart - 50th Percentile with Error Bars

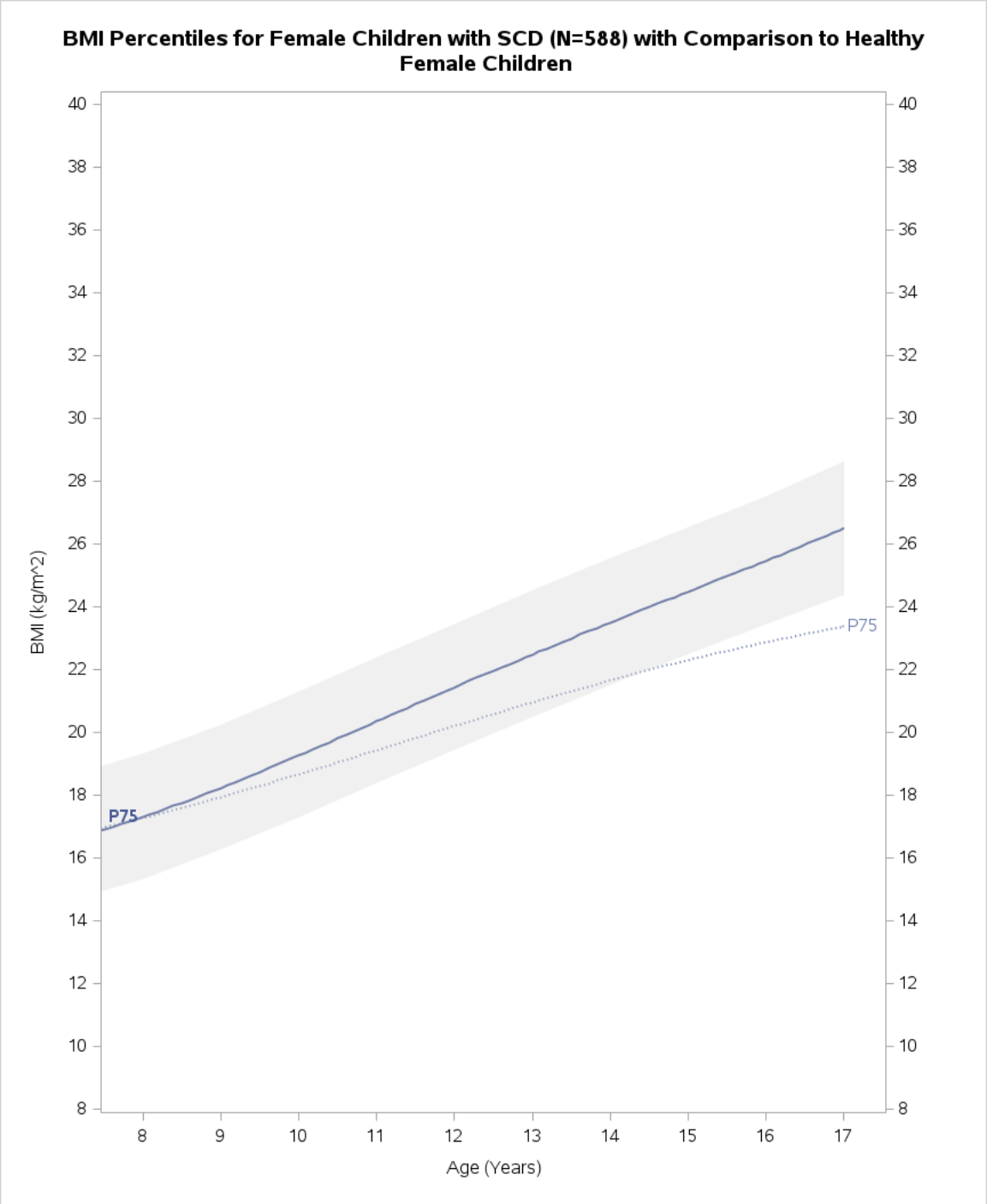

Figure A.16 – Female Growth Chart - 75th Percentile with Error Bars

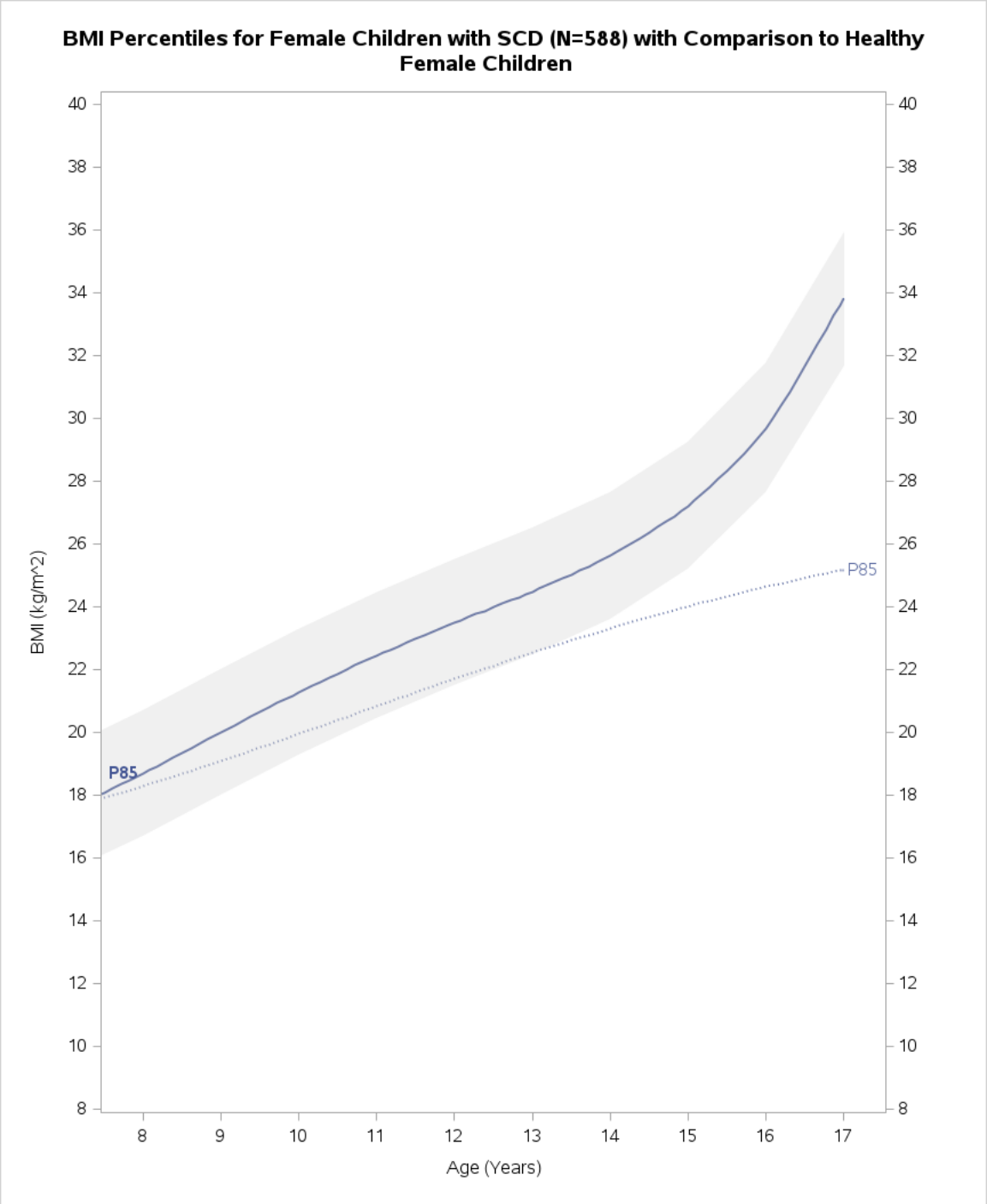

Figure A.17 – Female Growth Chart - 85th Percentile with Error Bars

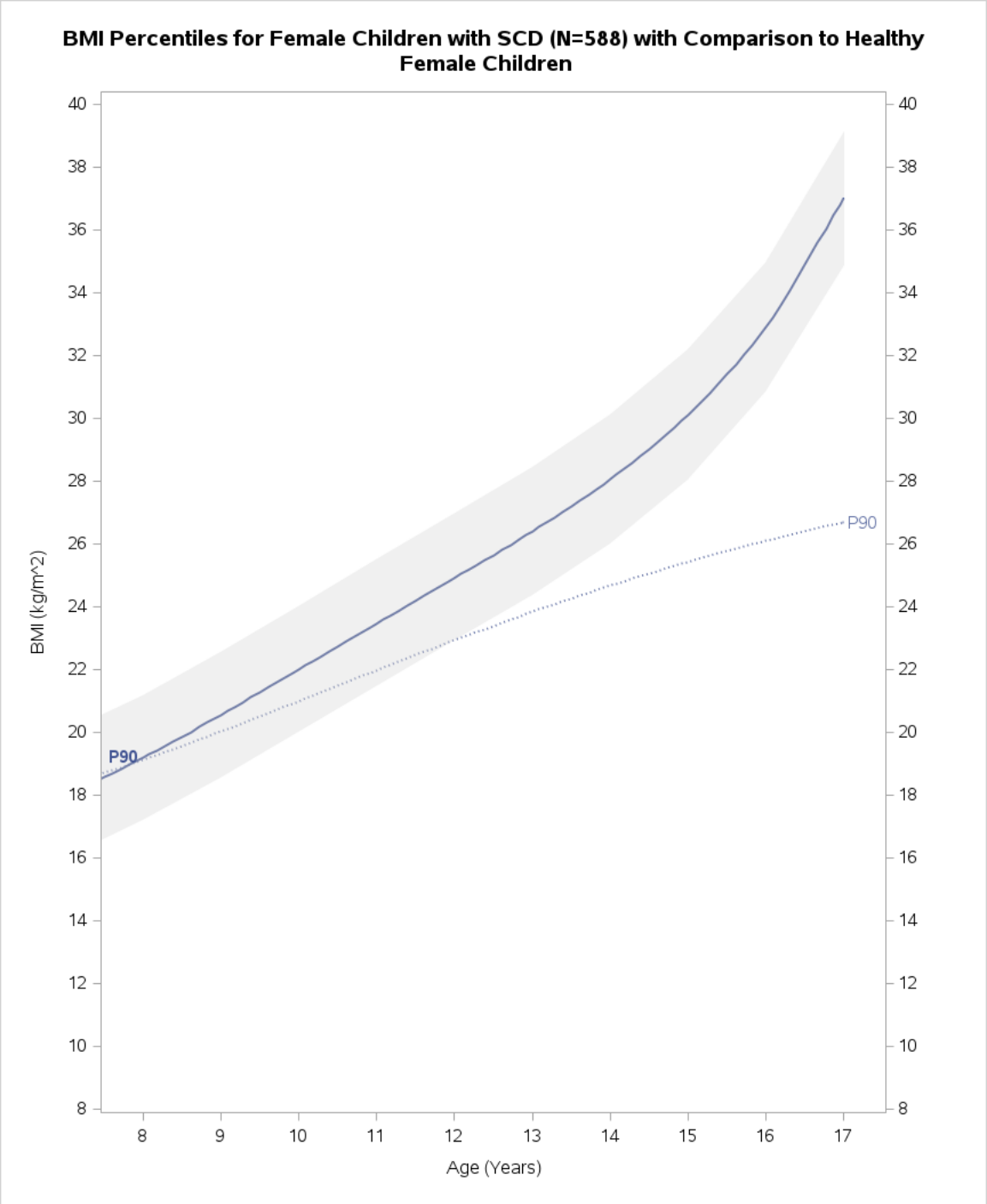

Figure A.18 – Female Growth Chart - 90th Percentile with Error Bars

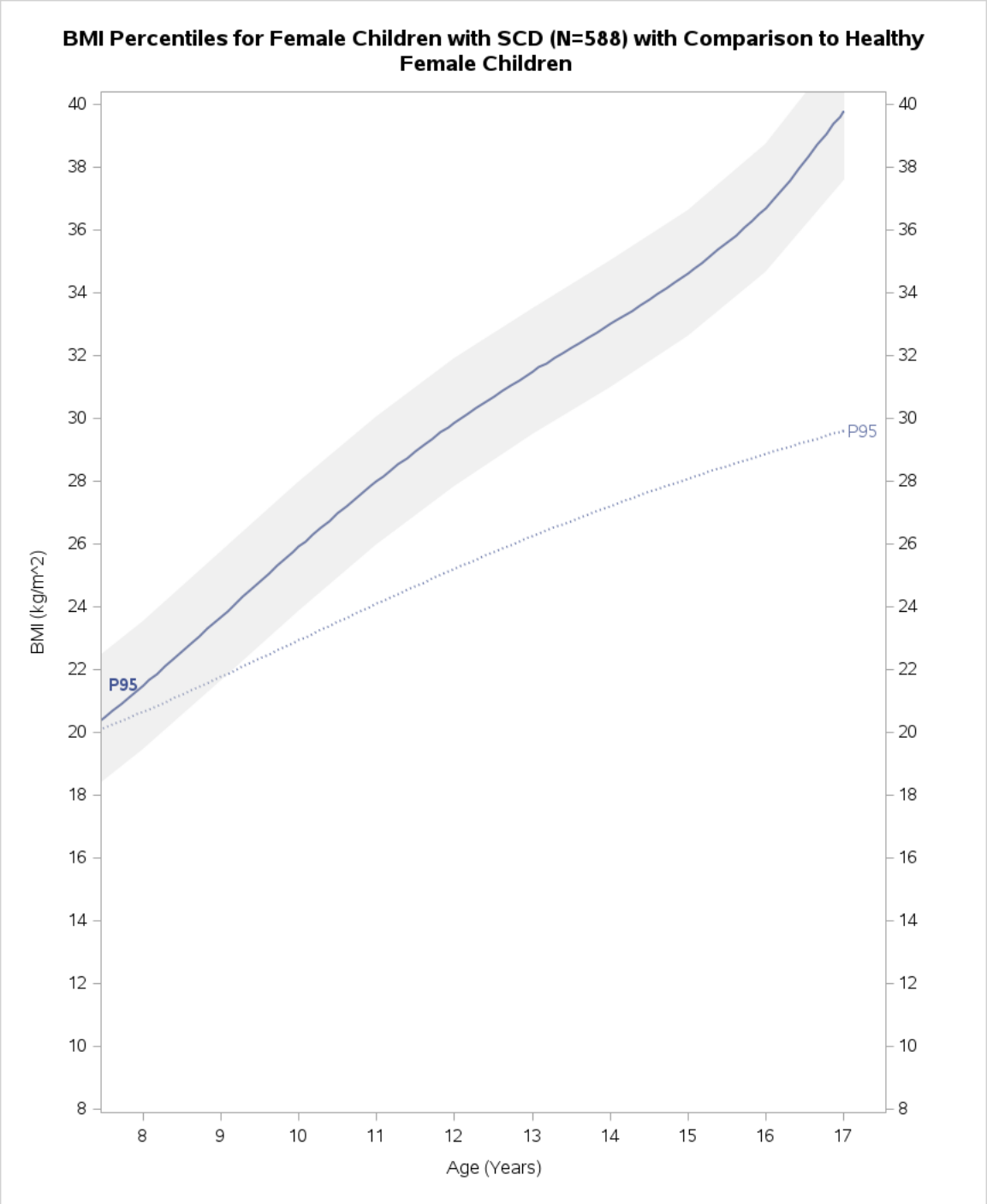

Figure A.19 – Female Growth Chart - 95th Percentile with Error Bars

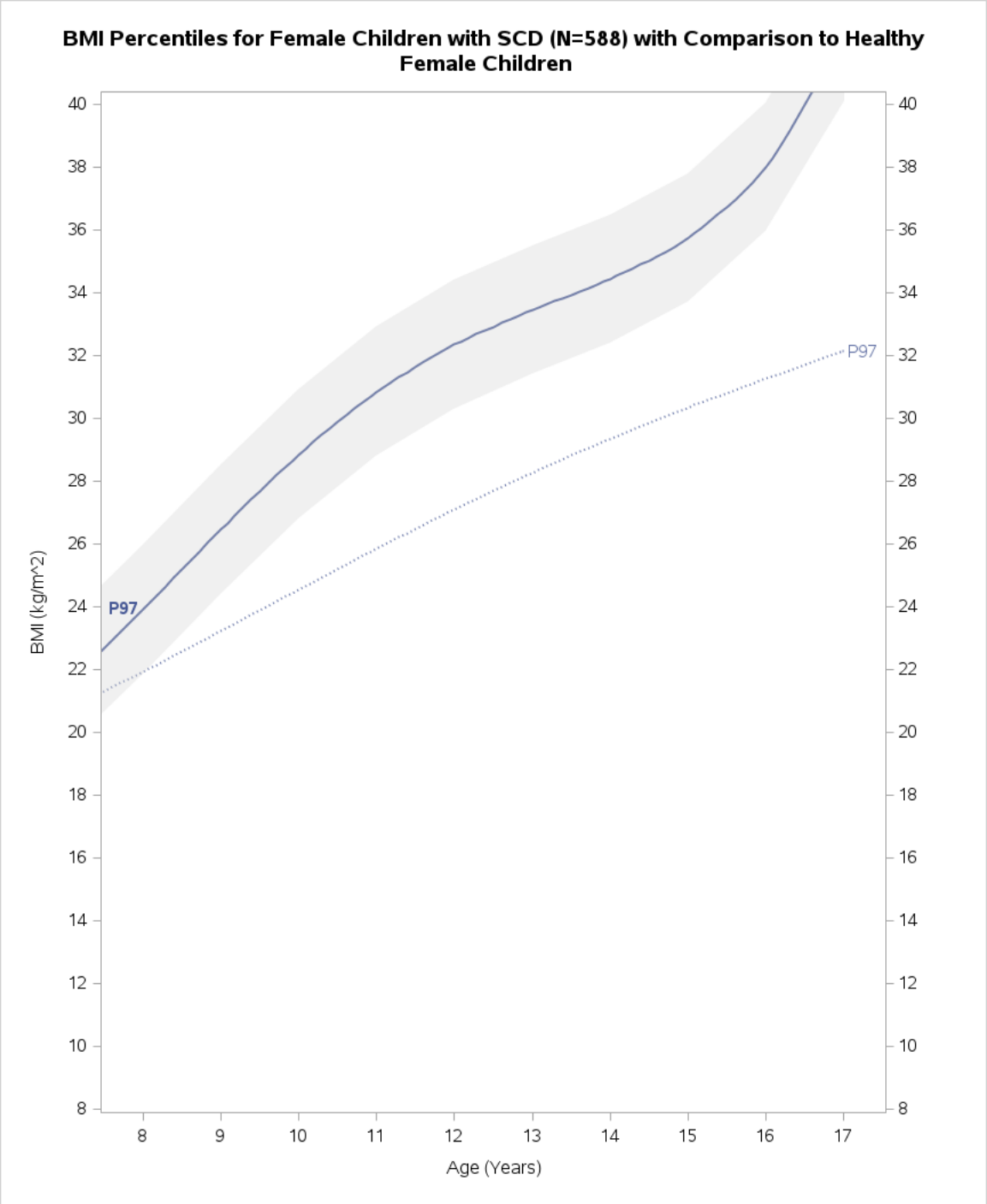

Figure A.20 – Female Growth Chart - 97th Percentile with Error Bars