Attrition in Weight Loss Programs

A Bayesian Statistics Approach for Predictive Insights

Maria Teresa Cardoso de Menezes

Dissertation written under the supervision of Professor Nicolò Bertani

Dissertation submitted in partial fulfilment of requirements for the MSc in Business Analytics, at the Universidade Católica Portuguesa
April 2024

[Page intentionally left blank.]

Attrition in Weight Loss Programs
A Bayesian Statistics Approach for Predictive Insights

Maria Teresa Cardoso de Menezes

April 2024

Supervisor: Professor Nicolò Bertani

Abstract

In today's service-oriented business landscape, client acquisition and retention are imperative, given the significant influence clients wield over service outcomes. This influence is particularly pronounced in prolonged services like weight loss programs, where client motivation towards program completion is crucial for achieving desired outcomes and ensuring customer satisfaction. This thesis investigates factors contributing to attrition in a three-phase weight loss program, leveraging a comprehensive dataset encompassing client registration details, progress tracking, and demographic information. However, the challenge of missing data arises when clients discontinue the program, hindering a comprehensive understanding of the weight loss journey. To address this challenge, a causal approach to missing data imputation is adopted, utilizing Bayesian Statistics to harness the inherent information within the data. Through extensive literature review and methodological exploration, the study sheds light on the reasons for attrition and proposes practical insights into addressing churn issues. Age and weight emerge as significant predictors of program completion, with older individuals exhibiting higher completion rates across all phases. Additionally, the study highlights the nuanced impact of previous program attempts on completion rates. Overall, this study contributes to the understanding of attrition in weight loss programs and offers valuable insights into leveraging Bayesian imputation models for predictive analytics in service-oriented contexts.

Keywords: Weigh Loss Programs, Attrition, Bayesian Statistics, Primates Imputation

[Page intentionally left blank.]

Attrition in Weight Loss Programs
A Bayesian Statistics Approach for Predictive Insights

Maria Teresa Cardoso de Menezes

April 2024

Supervisor: Professor Nicolò Bertani

Resumo

No atual panorama empresarial centrado no serviço, a aquisição e retenção de clientes são imperativas, dada a influência significativa que estes exercem sobre os resultados dos serviços. Esta influência é particularmente pronunciada em serviços prolongados, como programas de perda de peso, onde a motivação do cliente para a conclusão do programa é crucial para alcançar os resultados desejados e garantir a satisfação do cliente. Esta tese investiga os fatores que contribuem para a desistência num programa de perda de peso de três fases, utilizando um conjunto de dados que inclui detalhes de registo do cliente, monitorização do progresso e informações demográficas. No entanto, surge o desafio de falta de observações quando os clientes interrompem o programa, dificultando uma compreensão abrangente da jornada de perda de peso. Para enfrentar este desafio, é adotada uma abordagem causal para a imputação de dados em falta, utilizando Estatísticas Bayesianas para aproveitar a informação inerente nos dados. Este estudo apresenta razões para a desistência e propõe perspetivas práticas para enfrentar questões de rotatividade. A idade e o peso emergem como preditoes significativos da conclusão do programa, com os indivíduos mais velhos exibindo maiores taxas de conclusão em todas as fases. Além disso, o destaca o impacto subtil das tentativas anteriores de programa nas taxas de conclusão. No geral, este estudo contribui para a compreensão da desistência em programas de perda de peso e oferece insights valiosos sobre a utilização de modelos de imputação bayesianos para análises preditivas em contextos orientados para o serviço.

Keywords: Programas de Perda de Peso, Atrito, Estatística Bayesiana, Primates Imputation

[Page intentionally left blank.]

Acknowledgements

Firstly, I would like to express my heartfelt gratitude to Professor Nicolò Bertani for his invaluable guidance and unwavering support throughout this journey. Professor Bertani not only introduced me to the field of Bayesian Statistics but also provided me with expert knowledge, insightful recommendations, and endless patience. I am truly grateful for his mentorship and dedication.

Secondly, I am deeply thankful to Católica for providing me with the opportunity to study at such a prestigious Institution. I am also indebted to the exceptional professors who have imparted their knowledge and wisdom, enriching my academic experience.

I would also like to extend my appreciation to my friends for their unwavering support and encouragement, especially during those Sunday breakfasts where they assured me that I could persevere through this endeavor. A special acknowledgment goes to Tomás, who patiently listened to my complaints and bolstered my spirits, serving as a constant source of motivation.

Lastly, but certainly not least, I am profoundly grateful to my incredible family. They have been my pillars of strength, enduring countless conversations about Bayesian statistics and weight loss programs. Their unconditional love and support have empowered me to achieve everything that I have in my life. A heartfelt thanks goes to my grandmother, to whom I dedicate this work. She has always been my greatest believer, going above and beyond to ensure that I have every opportunity to excel not only in my education but in everything that I do.

[Page intentionally left blank.]

Contents

List of Tables

List of Figures

1 Introduction

In today's business landscape, clients play a crucial role in the delivery of services, making client acquisition and retention paramount for service providers. Specifically, customers have a significant impact on service outcomes across various sectors, shaping their own satisfaction and the derived value (Bitner et al., 1997). This influence is particularly evident in complex and prolonged services such as healthcare, weight loss programs, education, and sports (Buckinx and Van den Poel, 2010). In these contexts, active customer engagement is vital, as customers contribute diverse resources—cognitive, emotional, and physical inputs—within their respective domains. This collaborative effort spans extended periods, guided by established directives from service providers.

In the specific context of weight loss programs, the central focus of this thesis, it is imperative for companies providing such services to maintain a heightened awareness of the factors shaping client motivation towards program completion. It is crucial to identify elements that could potentially lead to a decrease in client motivation, resulting in user churn within these programs. This heightened awareness is of paramount importance for fostering and sustaining customer adherence, as it directly aligns with achieving customer goals, enhancing satisfaction levels, and eliciting positive word-of-mouth endorsements.

This thesis aims to investigate and identify factors contributing to attrition in a three-phase weight loss program. Specifically, the goal is to understand the reasons individuals do not complete this program. The program encompasses a comprehensive dataset, incorporating client registration details, progress tracking during consultations, as well as demographic, marketing, and clinical data. However, when a client discontinues the program, we lose access to their information, particularly their progress along the weight loss journey. This poses a challenge because, as mentioned earlier, prior research indicates that weight loss outcomes significantly influence the client's decision to continue in such programs. Since clients drop out, we are faced with a missing data issue.

To overcome this challenge, we approach the issue of missing data causally. In this context, any generative model inherently contains information about variables that haven't been observed. We aim to utilize this information for imputing missing values, employing Bayesian Statistics as our method of choice.

We commence by conducting a comprehensive literature review elucidating the reasons individuals tend to discontinue weight loss programs. Furthermore, we delve into the concepts of attrition and churn in these programs, expanding our exploration to other relevant domains. Subsequently, we delve into the literature surrounding missing data problems and elucidate

strategies for their resolution. Finally, we delve into the theoretical underpinnings of our chosen approach, utilizing Bayesian statistics. Moving forward, we will briefly contextualize the data and business environment, while defining attrition within the scope of our study. We will then undertake descriptive statistics, crucial for informing our subsequent modeling approach, ensuring optimal utilization of computational resources.

Our modeling approach is divided into two distinct phases: initially, an exhaustive exploration of the entire program, delving into factors correlated with attrition throughout the weight loss regimen; this is succeeded by an in-depth analysis of each phase individually. In the results section, we demonstrate the efficacy of this approach, highlighting the significant impact of weight, both initial and final, on program completion, spanning not only the entirety of the diet but also each distinct phase. Furthermore, our findings corroborate previous research, indicating age as a notable determinant of attrition in such programs, alongside the influence of the patient's chosen consumption method, be it loose consultations or specific packs integrating multiple consultations, whose impact varies across phases as anticipated. These findings are thoroughly examined and contextualized in the discussion section.

In conclusion, this thesis seeks to advance understanding of attrition in a three-phase weight loss program and offer practical insights into addressing churn issues through the lens of Bayesian Statistics.

2 Literature Review

In this section, we embark on a comprehensive literature review, beginning with an examination of factors contributing to attrition in weight loss programs. Subsequently, we delve into related studies exploring attrition not only in weight loss programs but also in healthcare services and other business domains. Our exploration extends to understanding how missing data is handled in programs akin to the one under study, as well as in other fields.

Following this, we provide a theoretical background on the techniques slated for addressing this challenge. These encompass Bayesian Statistics, Multilevel Models, Markov Chain Monte Carlo, Hamiltonian Monte Carlo, and imputation methods.

2.1 Why do people drop out from weight-loss programs?

In an increasingly health-conscious society, the popularity of weight loss programs has surged. However, despite their widespread adoption, many individuals discontinue these programs prematurely for various reasons. These may range from dissatisfaction with the progress of interventions to logistical challenges in attending scheduled sessions. This phenomenon,

commonly termed ”attrition”, poses a significant challenge to the effectiveness and sustainability of weight management programs (Fayyaz et al., 2022).

The existing body of literature has thoroughly delved into various factors that could potentially correlate with or predict attrition in weight loss programs. The main factors that are typically studied are demographic factors, psychological and physical health of the patient, health-related behaviours, personality traits, logistical considerations regarding the treatment, and weight and shape. Despite this extensive exploration, no consistent patterns or predictors of treatment attrition have emerged within these domains (Moroshko et al., 2011).

Nonetheless, some subtle trends can be identified. Age, for instance, appears to be the most consistently observed demographic predictor for attrition in weight loss programs. Studies such as Fabricatore et al. (2009) and Honas et al. (2003) have determined that older individuals exhibit a higher likelihood of successfully completing weight loss programs. This propensity may be attributed to the notion that older participants often possess more consistent and foreseeable responsibilities, aligned with heightened motivation to pursue weight loss goals driven by increased health concerns or limitations in quality of life. Further research, as exemplified by Teixeira et al. (2004), indicates that individuals facing pronounced mental health issues, such as depression, are also less inclined to successfully conclude weight loss challenges. The requirements of a weight loss program can be particularly daunting for participants experiencing heightened depressive symptoms, as these individuals often contend with diminished energy, motivation, and concentration associated with this condition. Finally, as Moroshko et al. (2011) notes, practical obstacles like the distance to the consultation location, financial constraints, along with clients' elevated weight loss expectations or encountering lower initial weight loss, can lead to attrition.

2.2 Predicting Attrition

Much of the existing research on weight-loss interventions predominantly focuses on obesity patients, emphasizing weight loss outcomes and the resolution of obesity-related comorbidities, often neglecting to report attrition rates and predictors (Moroshko et al., 2011). The prediction of attrition in weight management programs has been investigated across various contexts, with a predominant reliance on conventional methodologies like linear and logistic regression (Fabricatore et al., 2009; Honas et al., 2003). In a contemporary vein, recent investigations have embraced cutting-edge techniques, incorporating neural networks to enhance the precision of attrition prediction, exemplified in the work of Fayyaz et al. (2022).

Additionally, patient retention and attrition pose pervasive challenges in healthcare programs and clinical trials (Kearney et al., 2018), prompting extensive research in this area.

Within the current literature, diverse studies explore these challenges in various settings. Some concentrate on predicting patient attendance, attempting to forecast whether patients will attend scheduled appointments. Noteworthy examples include studies specific to certain specialties (Bush et al., 2014; Hayton et al., 2013) and those encompassing all appointments within hospitals or general practices (Dashtban and Li, 2022). Logistic regression also emerges as a commonly employed technique for predicting missing attendance (Carreras-García et al., 2020), although more recent approaches incorporate deep learning models for enhanced predictive accuracy (Dashtban and Li, 2022). Beyond attendance prediction, research in healthcare also addresses patient adherence, namely treatment and medication adherence (Kumamaru et al., 2018).

Furthermore, relevant studies outside the healthcare domain, such as churn prediction in economic contexts, seek to forecast engagement patterns of individuals, such as customers and employees, to optimize retention strategies (Tamaddoni Jahromi et al., 2014). These diverse strands of research contribute valuable insights to understanding and mitigating challenges associated with patient retention in healthcare interventions.

2.3 Missing Data Mechanism

One of the primary challenges in this research lies in the absence of data related to the progression of patients who drop out of the weight loss treatment program, as such information does not exist. Consequently, for all patients who discontinue, we encounter missing values for variables such as weight.

2.3.1 Missing Data in Weight Loss Programs

Many studies in the weight loss field overlook the treatment of missing data. Those that do address this issue typically resort to modified versions of the last observations, employing average approximations from the last months or adopting averages from comparable studies (Teixeira et al., 2004). On the other side, when it comes to predicting attrition in other fields, we often observe a more frequent acknowledgment and resolution of missing data challenges. This is commonly managed within statistical analyses through either excluding missing data from the analysis population (complete case analysis) or employing imputation methods to estimate outcomes (Kearney et al., 2018).

In literature, mean-value imputation stands out as the most basic method for addressing missing data, involving the assignment of the mean value of a given variable to all instances with missing values (Rubin, 2018). Despite its simplicity in concept, this approach comes with significant drawbacks. One major issue is the reduction of sampling variability in the data, as all individuals with missing values for a specific variable end up being imputed with the identical

value. Additionally, mean-value imputation overlooks the intricate multivariate structure of the data (Austin and Escobar, 2005). For example, if data tend to be missing for bulkier patients, this method inadvertently distorts the representation of weightier individuals by imputing mean values from patients with complete data, who, in contrast, tend to be lighter. Moving forward, to address the challenge of missing data, we will delve into the causative factors behind the missing values and adopt a Bayesian inference approach to tackle the issue of attrition.

2.3.2 Complete Case Analysis

Robins et al. (1994) propose that when the probability of missing data is solely contingent on completely observed predictor variables, a complete case analysis will yield regression coefficient estimates that are asymptotically unbiased. However, inconsistency in estimation may arise when the likelihood of missing data depends on both the outcome variable and the fully observed variables. Therefore, within the context of this thesis, opting for a complete case analysis is not justified. This approach would entail discarding valuable data and likely introducing bias into our inferences due to the strong correlation between missingness and the outcome. The absence of data is not arbitrary; there is a clear reason for not observing these variables – the client has dropped out. Consequently, it is imperative to acknowledge the underlying process causing this missing data and handle these cases with special attention rather than straightforward exclusion.

2.3.3 Censored data as missing data

In our case study, the probability of final weight being missing depends on various factors and, notably, on the value of the final weight itself. Consequently, the data are not observed at random. Furthermore, in this scenario, the missing data mechanism is considered non-ignorable. Despite this, since we have information about the missing-data mechanism, it is still possible to conduct an appropriate analysis (Gelman, 2013).

The realm of missing data encompasses various mechanisms, including instances of data missing completely at random, censoring, and truncated data. However, the focus of our inquiry predominantly centers on addressing issues associated with censoring within our dataset. Censoring arises when the event of interest fails to manifest within the established observation window. Such circumstances may occur due to the termination of observation before the event unfolds, or due to intervening factors that render the event unobservable . For instance, a patient may discontinue treatment before the final weight measurement can be obtained.

In essence, censored values represent instances where the true final weight remains undisclosed — a form of missing data (McElreath, 2023). Despite this absence, pertinent infor-

mation, such as the initial weight, often remains available, facilitating the narrowing down of plausible value ranges. Consequently, the formulation of a model capable of directly inferring each unobserved final weight becomes feasible.

In many empirical contexts, a comprehensive understanding of the distribution of observed data is attainable. Nevertheless, these datasets frequently exhibit censoring phenomena, wherein unobserved observations are deemed missing due to their occurrence beyond a predetermined threshold (Gelman, 2013). For instance, final weight measurements beyond 200kg may not be recorded. Within the scope of our investigation, while assumptions regarding the distribution of complete data for final weight are permissible, it remains imperative to acknowledge that the observed data are subject to censoring at an undetermined threshold, denoted herein as ϕ.

2.4 Bayesian Data Analysis

Bayesian Data Analysis, rooted in Bayes' theorem, offers a distinctive approach to ordinary data analysis and parameter estimation. Unlike conventional statistical methods, Bayesian statistics considers both observed and unobserved parameters within a statistical model by assigning a joint probability distribution, incorporating prior and data distributions.

The Bayesian workflow typically involves three key steps: first, encoding existing knowledge about parameters through the prior distribution, often established before data collection; second, constructing the likelihood function based on available parameter information from the observed data; and finally, combining the prior distribution and likelihood function using Bayes' theorem to compute the posterior distribution. This posterior distribution encapsulates updated knowledge, striking a balance between prior beliefs and observed data, facilitating meaningful inferences. Bayesian inferences achieve optimality when averaged across this joint probability distribution, relying on their conditional distribution given the observed data (van de Schoot et al., 2021).

2.4.1 Bayesian Inference

As denoted by Gelman (2013), the essence of Bayesian statistical inference lies in making probability statements about parameters (θ) or unobserved data ($\widetilde{y}$) in terms of conditional probabilities, denoted as $p(\theta|y)$ or $p(\widetilde{y}|y)$, respectively. These statements are based on conditioning on the observed data (y) and any known covariates (x), which distinguishes Bayesian inference from traditional statistical methods that typically evaluate estimators retrospectively.

Bayesian inference begins with a joint probability distribution for θ and y, expressed as the product of the prior distribution $p(\theta)$ and the sampling distribution (or data distribution) $p(y|\theta)$,

yielding the joint probability mass or density function

$$p(\theta, y) = p(\theta)p(y|\theta) \tag{1}$$

.

Applying Bayes' rule to condition on the observed data y results in the posterior density

$$p(\theta|y) = \frac{p(\theta, y)}{p(y)} = \frac{p(\theta)p(y|\theta)}{p(y)}, \tag{2}$$

where $p(y)$ is the marginal likelihood, obtained by integrating out θ.

Predictive inferences, pertaining to unknown observable ($\widetilde{y}$), follow a similar logic. Before considering the data y, the distribution of $\widetilde{y}$ is defined as the prior predictive distribution, which encompasses all possible outcomes of $\widetilde{y}$ given θ and is not conditional on previous observations. After observing y, the posterior predictive distribution for $\widetilde{y}$ is derived, conditioning on the observed data, y.

Basically, Bayesian inference involves updating beliefs about parameters or unobserved data based on observed data, through the calculation of posterior distributions. Predictive inferences extend this framework to make predictions about future observations.

Additionally, in Bayesian modeling, the model functions as a processing engine, operating on data to derive a posterior distribution by conditioning the prior on observed data, following principles of probability theory (McElreath, 2023). However, formal conditioning can be impractical for complex models in contemporary science. To address this, conditioning engines capable of accommodating any useful prior for inference, such as Markov chain Monte Carlo (MCMC), are utilized, particularly in the context of multilevel models.

2.4.2 Multilevel Models

While multilevel models (MLMs) are not new, they have garnered attention among researchers due to advancements in computational capabilities. They complement Bayesian analysis effectively, offering several advantages. Bayesian MLMs mitigate issues of underfitting or overfitting commonly observed in single-level models, especially when employing repeated sampling. They handle uncertainty across uneven sample sizes automatically, particularly in within-subject designs. Additionally, Bayesian MLMs explicitly model variation within and between clusters of data, preserving uncertainty and minimizing the need for extensive data transformation (Felderer and Travassos, 2020).

In our specific context, where models for missing data (imputation) are crucial, Bayes' theo-

rem serves as the foundation for conducting inference with MLMs. Markov chain Monte Carlo (MCMC) techniques drive the Bayesian analysis, providing a robust computational engine for exploring parameter spaces and deriving posterior distributions.

2.4.3 Markov Chain Monte Carlo

Markov chain simulation, also known as Markov chain Monte Carlo (MCMC), stands as a versatile method for approximating posterior distributions in Bayesian statistics. This technique operates by iteratively drawing values of θ from approximate distributions and subsequently refining those draws to better reflect the target posterior distribution, $p(\theta|y)$. The sequential nature of this sampling process imbues it with a Markovian property: each draw's distribution depends solely on the previously sampled value, resulting in a Markov chain. In probability theory, a Markov chain is formally defined as a sequence of random variables, $\theta_1, \theta_2, ...$, where the distribution of θ given all preceding θ's is contingent solely on the most recent value, θ_{t-1}. (Gelman, 2013)

However, the efficacy of MCMC hinges not solely on the Markovian property, but rather on the iterative improvement of approximate distributions throughout the simulation, ultimately converging towards the target distribution. This iterative refinement process ensures that the sampled values increasingly approximate the desired posterior distribution, enhancing the method's effectiveness (Gelman, 2013).

Markov Chain Monte Carlo (MCMC) remains a cornerstone technique in Bayesian statistics, furnishing a potent means to generate samples from intricate posterior distributions. Rooted in the principles of Markov chains, MCMC proves invaluable for modeling dynamic systems across a wide array of (McElreath, 2023). Figure 1 represents an example of a Markov chain simulation with a Metropolis algorithm, where θ is a vector with only two components, with a bivariate unit normal posterior distribution, $\theta \sim N(0, I)$.

Notable MCMC algorithms, such as Metropolis Algorithm (illustrated above), Gibbs Sampler, and Hamiltonian Monte Carlo (HMC), play a pivotal role in Bayesian inference, facilitating the exploration of posterior distributions encapsulating updated probabilities of parameters given observed data and prior beliefs (McElreath, 2023).

2.4.4 Hamiltonian Monte Carlo

Hamiltonian Monte Carlo (HMC) stands out as an advanced MCMC algorithm designed to enhance sampling efficiency, especially in exploring high-dimensional and complex probability distributions. It addresses common limitations encountered in traditional MCMC methods, such as slow parameter space exploration and high autocorrelation between samples.

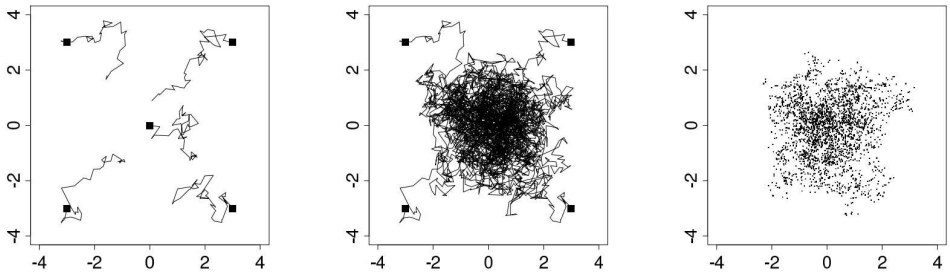


Figure 1: Markov Chain Simulation (Gelman, 2013)

Note: The graphs displays various stages of a simple Markov simulation. The first graph (left) illustrates the initial phases of the Markov chain simulation, while the second graph (middle) represents the mature stage of the simulation. The third graph (right) displays the iterates from the second halves of the sequences, comprising a collection of correlated draws from the target distribution.

Although HMC entails higher computational costs compared to methods like Metropolis or Gibbs sampling, its proposals exhibit superior efficiency. Consequently, HMC requires fewer samples to characterize the posterior distribution effectively. This efficiency becomes increasingly pronounced in scenarios involving complex models with numerous parameters. Conceptually, HMC simulates a physics-inspired process where a frictionless particle traverses a surface defined by the logarithm of the posterior distribution. The efficiency of HMC proposals hinges on intelligent path selection, guided by the gradient of the log-posterior. While HMC aims for acceptance of every proposal, it employs a rejection criterion to account for numerical approximations (McElreath, 2023).

Despite its inherent complexity, HMC offers efficient sampling of intricate models where conventional methods struggle to navigate parameter space effectively. The integration of HMC into Bayesian workflows, facilitated by tools like the `ulam` function from the `rethinking` package in R, streamlines model preparation and posterior sampling.

Posterior sampling is executed with the aid of HMC, leveraging the computational prowess of `stan`. During this phase, multiple chains are concurrently executed, each undergoing a warmup period. The warmup phase, akin to tuning an instrument, adjusts parameters such as step size and leapfrog steps to optimize exploration of the posterior landscape. Once the warmup phase concludes, the chains proceed with actual sampling from the posterior distribution. This process yields a collection of samples representative of the posterior distribution, providing insights into parameter estimates and uncertainty quantification.

2.4.5 Imputing Primates

This thesis adopts the methodology showcased by McElreath (2023), which leverages Bayesian imputation to handle missing values in variables. In this approach, each missing value is associated with a unique parameter, while observed values contribute to the overall understanding of the distribution of the variable. This distribution serves as a prior for the missing values and is further refined within the comprehensive model. Consequently, a posterior distribution is derived for each missing value, providing updated and more informed estimations.

Implementing an imputation model involves various methods, each with its challenges, particularly in managing indices to respect the locations of missing values. To address this, we adopt a method that merges observed values and parameters into a vector, treating them collectively as "data" in the regression process. This merging process can be conveniently automated using the `ulam` function from the `rethinking` package.

Furthermore, for the Bayesian inference process, we employ the `ulam` function, which streamlines Stan code — a standalone programming language used for defining and estimating Bayesian models. This facilitates the utilization of Hamiltonian Monte Carlo (HMC) for posterior sampling. By adopting this approach, we simplify the implementation of imputation models and ensure efficient handling of missing values within the Bayesian framework.

3 Business Context

This thesis is the result of collaboration with a Portuguese company specializing in the development and sale of nutrition services, including a three-phase weight loss program, and a range of food supplements tailored to support weight loss.

From a business perspective, client attrition presents significant challenges. Not only does it impede customers from achieving their desired goals, thereby potentially diminishing satisfaction levels and limiting positive word-of-mouth endorsements, but it also impacts the company's revenue stream. The primary source of revenue for the company is derived from the sale of supplements and healthy food within its weight loss program brand. These products are specifically recommended by nutritionists based on each patient's individual circumstances, progress, and dietary phase. Consequently, it is anticipated that clients who demonstrate a higher level of commitment to the program by diligently following each step are not only likely to achieve better results but also invest more in the company's products.

Understanding the factors contributing to client dropout is crucial from various perspectives, business, nutritional, and marketing. By pinpointing these factors, the company can

strategize to mitigate them effectively. This entails not only preserving client engagement and satisfaction but also optimizing revenue generation opportunities through targeted interventions and tailored offerings.

3.1 Weight Loss Program

The company offers a personalized weight loss program known as the ”3 Phase Method,” which focuses on individual dietary preferences to promote sustainable dietary re-education. This method is supported by a specialized team of nutritionists who provide weekly consultations, incorporating cognitive-behavioral strategies and applied coaching techniques to enhance treatment adherence and achieve better weight reduction and maintenance results.

The 3 Phase Method comprises the following phases:

- **Phase 1:** First Lasting approximately three weeks on average, this phase aims to stimulate catabolic pathways to promote lipid oxidation. It involves reducing carbohydrate intake to 30-50 g/day to decrease postprandial insulin secretion and plasma glucose levels, thereby reducing hunger sensations, particularly in individuals with insulin resistance. Protein intake is increased, as it is the most satiating and thermogenic macronutrient, facilitating weight loss. High-protein diets offer advantages in maintaining fat-free mass in an energy-restricted context.

- **Phase 2:** This phase involves a gradual increase in carbohydrate intake while continuing to reduce weight and body fat. Its duration varies depending on the individual’s excess weight and therapeutic adherence. When the target weight is achieved, the third phase begins.

- **Phase 3:** In this phase, consultation frequency becomes less frequent as therapeutic progress is made and healthy eating habits are established. Ideally, this phase includes a one-year follow-up (approximately eight consultations), aligning with international guidelines for weight maintenance. The main goals of this phase are weight management through healthy, varied, and balanced eating, enhancing nutritional literacy, and improving self-control and self-regulation abilities.

Overall, the this weight loss program aims to provide personalized nutritional intervention tailored to the needs of individuals with excess weight, emphasizing long-term behavior change and sustainable weight management.

3.2 Defining Attrition

As previously mentioned, attrition in weight loss programs signifies the premature discontinuation of these programs by individuals. Our primary objective is to predict attrition within a 3-phase weight loss program. To define attrition, we establish completion criteria for the entire program, as well as for each individual phase.

In practical terms, completing the full diet program entails progressing through all three phases, which includes attending at least 8 consultations in phase 3. Phase 3 serves as a maintenance phase, where clients are expected to demonstrate learned behaviors independently. While clients may continue attending sessions beyond the minimum requirement, they are considered to have completed the program if they meet this criterion. To operationalize attrition, we introduce a binary variable called `completed_diet`, which takes a value of 1 when no attrition occurs, signifying program completion, and 0 when attrition is observed, indicating non-completion.

Given the nature of our data, we extend our analysis to examine attrition within each specific phase of the program. For phase 3, completion is straightforward, as it requires a fixed minimum of 8 consultations. Hence, `completed_phase_3` is set to 1 if the patient has completed at least 8 consultations in this phase, and 0 otherwise. However, measuring completion for phases 1 and 2 is more nuanced. Since there isn't a predefined minimum number of consultations for these phases, completion depends on the client's performance, commitment, and progress, evaluated weekly by the accompanying nutritionist. To determine completion status for phase 1, we consider any consultation in phase 2 as indicative of phase 1 completion, as the decision to transition to phase 2 implies successful completion of phase 1. Similarly, for phase 2, `completed_phase_2` is set to 1 if the patient has at least one consultation in phase 3, indicating completion of phase 2.

4 Data and Descriptive Statistics

4.1 Data Understanding

The dataset, covering the period from January 2016 to November 2023, meticulously documents the progress of 44,809 distinct patients, aged between 16 and 84 years old, on their weight loss journeys. Each entry in the dataset corresponds to a consultation session. This extensive dataset offers detailed insights at both the patient and consultation levels. At the patient level, it includes demographic information such as age, gender, and location. At the consultation level, it encompasses details like consultation date, location, and the attending nutritionist. Moreover, it contains vital clinical data, including the patient's weight, BMI, fat mass, and

more, recorded during each consultation.

The data is gathered by nutritionists who accompany patients during their consultations, systematically inputting it into the company's database. Demographic data is collected during the initial consultation for each patient, while consultation and clinical data are recorded during subsequent consultations. Consequently, we possess demographic data alongside consultation and clinical data for the initial consultation of every client. However, complete consultation and clinical data for every patient throughout their entire plan duration are not available due to attrition. When patients discontinue their participation (`completed_diet` == 0), data collection ceases. To address this, we've introduced an additional row for such cases. This row retains constant demographic data but contains missing values for variables not measured due to the patient's dropout.

4.2 Data Preparation

The dataset underwent significant data cleaning, particularly concerning outliers. Among the numerical variables, `atual_weight` presented the most prominent cleaning challenge. Instances where individuals purportedly weighed over 300 kilograms raised flags, as such weight not only seemed improbable but also practically unmeasurable given standard clinic equipment limitations, typically capped at around 200 kilograms.

Upon careful analysis and comparison of outlier observations with subsequent records for the same patients, it became apparent that most cases stemmed from typographical errors. Consequently, we adopted a systematic approach to address these anomalies. Observations falling within the range of 200 to 300 kilograms underwent individual scrutiny. For instances where weight exceeded 300 kilograms but remained below 2000 kilograms, we divided the weight by 10 to rectify potential data entry errors. Similarly, for patients with weights exceeding 2000 kilograms, we divided the recorded weight by 100 to bring it within the realm of plausibility. This method ensured data integrity while correcting for extreme outliers in weight measurements.

Regarding categorical columns, the `occupation` and `location` fields underwent extensive cleaning efforts due to their nature as open fields. This openness resulted in numerous inconsistencies and spelling errors. For the `occupation` variable, initially an open field for nutritionists to input until 2022, it later transitioned to a closed field named `profession`. Despite this change, the dataset still contained a staggering 4727 unique occupations. While most entries were successfully mapped to the occupation field, remaining cases underwent individual analysis and treatment. As for the `location` variable, consistently an open field, it contained a daunting 7,368 different postal locations, rendering manual resolution impractical.

To address this, we leveraged the postal code column and introduced a new column named "Location." This new column was populated using official postal code data obtained from the mail authority in Portugal[1]. By cross-referencing postal codes with this authoritative data, we were able to identify the corresponding district and municipality for each location.

In certain instances, patients chose to restart the program after dropping out, or they returned after some time following completion of the process. In such scenarios, we treated these individuals as new patients to ensure accurate monitoring of their progress from the outset. To streamline this monitoring process, we assigned them a new `patient_id`. Furthermore, we incorporated a new column to indicate the frequency with which each patient had repeated the diet. This addition offers valuable insights into their journey over time, allowing for a comprehensive understanding of their progress and engagement with the program.

4.2.1 Stratified Sample

As previously mentioned, our study involves not only employing Bayesian statistical inference to make probability statements about parameters predicting attrition but also addressing unobserved data, particularly final weight, which demands considerable computational time and resources. Due to constraints in these aspects, we were unable to utilize the entire dataset in our models. To circumvent these purely operational constraints, we opted to create random samples comprising 3000 unique patient IDs, along with all corresponding observations, from our dataset.

We made concerted efforts to ensure that our samples were as representative as possible of our entire dataset. To achieve this, we employed stratified sampling, wherein we divided subjects into two strata based on whether they completed the diet or not. Subsequently, we randomly sampled individuals from each stratum while maintaining the completion ratio observed in the original dataset. For example, considering a completion rate of approximately 10% for the entire diet program (as elaborated in the descriptive statistics section), we ensured that our stratified sample included 10% of patients who completed the diet and 90% who did not.

Furthermore, to conduct analyses for each phase individually, we created separate random samples comprising 3000 patients from each phase, while still adhering to the respective completion percentages observed in each phase. This approach allows for a comprehensive exploration of attrition dynamics across different phases of the diet program. Further details regarding this strategy and its implications will be elucidated in subsequent sections.

[1]The data was retrieved from the **official CTT website** on November 5, 2023.

4.3 Descriptive Statistics

Once again, Bayesian imputation, owing to its computational complexity, necessitates significant processing time, particularly given the extensive nature of our dataset comprising numerous variables. Consequently, prioritizing descriptive statistics becomes paramount. Hence, comprehending the nuances of our data empowers us to make informed decisions regarding resource allocation and helps mitigate unnecessary time and computational resource consumption during code execution.

In this section, we delve into the descriptive statistics of the entire dataset, rather than solely focusing on random samples. This approach enables us to gain a comprehensive understanding of the entire diet program. Importantly, we aimed to preserve the inherent characteristics of the entire dataset within the random sample, ensuring its representativeness.

Out of the 44,809 patients aged between 16 and 84 enrolled in our program, a mere 4,506 individuals have successfully completed the entire regimen. This stark statistic translates to roughly 10.01% of the total patient population, highlighting a significant prevalence of attrition within the diet program. Figure 2 illustrates this declining trend by showcasing the total number of patients participating in each consultation. Despite not having a fixed number of consultations for phases 1 and 2, we observe a distinct drop-off, particularly noticeable between the initial consultation, where all 44,809 patients are engaged, and the 10^{th} consultation, where only 22,088 participants remain. This suggests that a substantial portion of participants, roughly 50%, discontinue the program between phase 1 and phase 2.

Upon examining Figure 3, we gain valuable insights into the progression within each phase of the program. Notably, in Figure 3(a) we see that only approximately 25% of patients drop out before completing the first phase. This dropout rate appears reasonable, given that the first phase is typically the shortest, lasting an average of 3 to 4 consultations. Participants are still at the early stages of the program and tend to be highly motivated.

However, Figure 3(b) presents a different trend for Phase 2, where more patients churn (a total of 19,951) compared to those completing it (13,832), representing only a 41% completion rate. This shift can be attributed to the expectations of Phase 2, where significant weight loss is anticipated. Patients may become demotivated if they fail to observe desired results or, conversely, may feel they have achieved their goals and opt to discontinue the program.

Furthermore, this trend persists in Figure 3(c), with more individuals dropping out than completing Phase 3. Only 4,506 patients complete Phase 3 and the entire diet program, accounting for roughly 33% of Phase 3 participants and a mere 10.01% of overall diet partici-

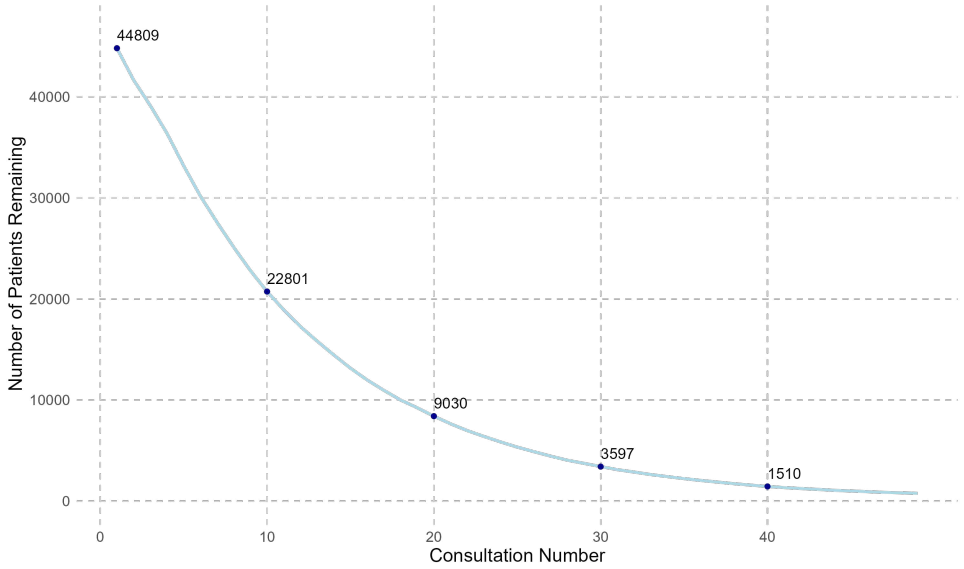


Figure 2: Number of patients over time

Note: The graph displays the progression of remaining patients throughout the consultation period. "Remaining patients," or "net patients," refer to the individuals continuing within the program. This value is derived by subtracting patients who did not proceed beyond the last consultation from the total number of patients in the previous consultation.

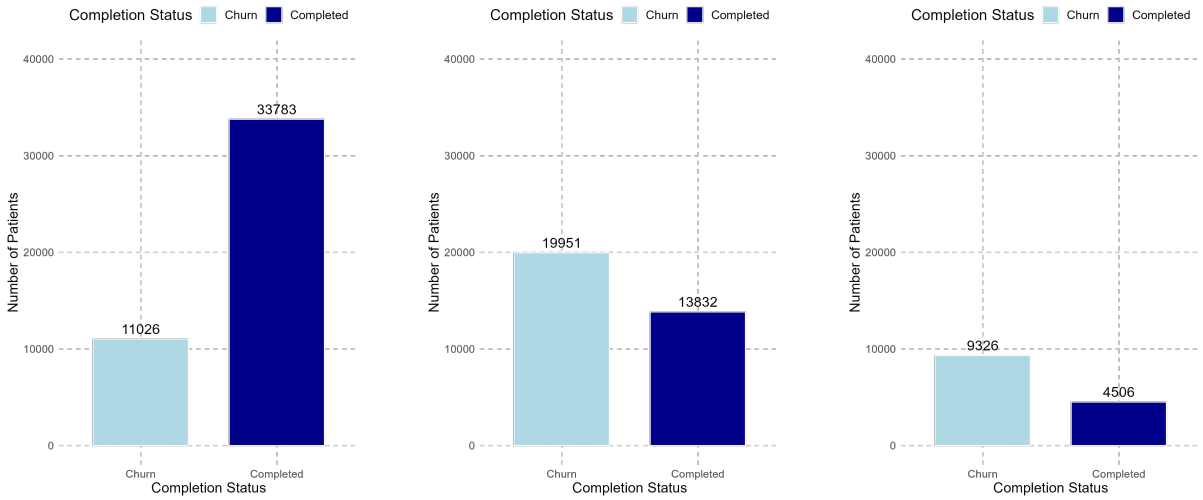
pants.

To deepen our understanding of the data, we conducted an analysis of relative frequencies between certain variables (those without missing values) and the program outcome. In general, we observed minimal differences between categories, with variances typically hovering around 2 percentage points. However, one variable stands out: age. Figure 4 illustrates the percentage of individuals who completed the diet versus those who churned, segmented by age group.

As anticipated, the data suggests that older individuals are more inclined to complete the diet. This observation will be subjected to rigorous testing in our subsequent models. Additionally, variables such as `marketing_data`, which denotes the source through which participants learned about the diet, and `brand_adhesion_type`, indicating how participants acquired the diet (either through loose consultations or specific packs), also displayed variations across categories. These variables warrant further investigation and are deemed pertinent for inclusion in our modeling efforts.

5 Methodology

As previously mentioned, the primary aim of this thesis is to identify the factors influencing whether individuals adhere to a diet regimen or not. Traditional prediction methods are inadequate due to missing values in crucial variables, such as `final_weight`, stemming from



(a) Phase 1 (b) Phase 2 (c) Phase 3

Figure 3: Completion Status in each Phase

Note: The graphs represent the number of individuals who churned and completed each phase. Only individuals who started each phase were considered; for instance, for a patient to be categorized as churned in phase 3, they must have completed phase 2. This approach ensures accurate reflection of progression and attrition within each phase.

participant dropouts. Employing Bayesian Imputation Models entails significant computational resources and time, particularly with a dataset as substantial as ours, comprising 44,809 distinct patients, many of whom have undergone multiple consultations (some exceeding 200). Given the impracticality of modeling with the entire sample, we opted to utilize random samples of 1000 patients each (with corresponding observations) for our analysis. This decision, while acknowledging potential risks to statistical power and precision, was made to strike a balance between computational feasibility and analytical robustness.

Our analysis is bifurcated into two parts. Initially, we investigate attrition within the entire diet program—what factors contribute to client churn? Subsequently, we delve into individual phases of the diet regimen for a more nuanced examination.

5.1 Complete Diet

5.1.1 Weight

We initiated our analysis with the most straightforward model, focusing on weight as a fundamental aspect of diet adherence. According to the literature review, weight serves as a significant motivator for individuals engaging in diet programs. We selected two distinct variables to capture this aspect: the `initial_weight`, representing the first measurement recorded for each patient, and the `final_weight`, which necessitated imputation to address missing values.

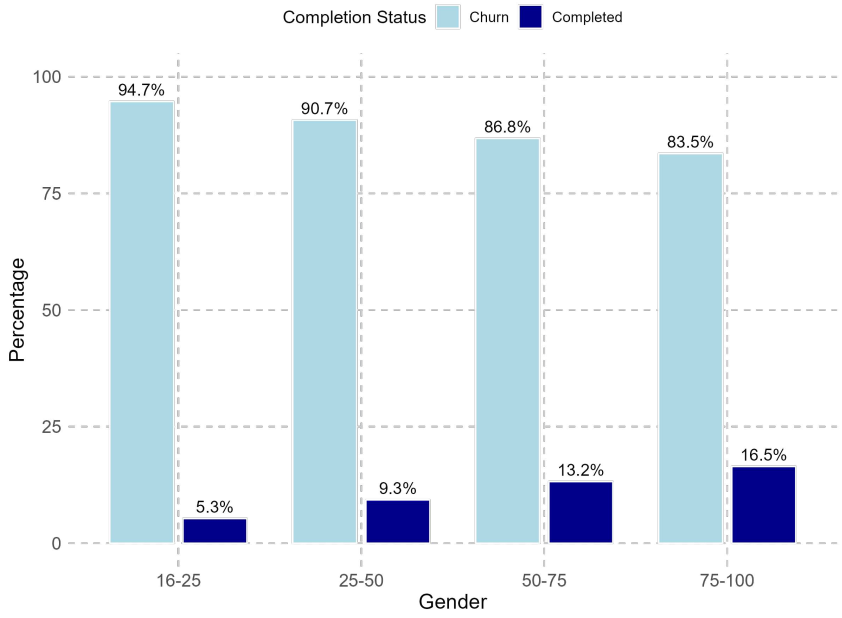


Figure 4: Distribution of Completion Status by Age Group

In practical terms, our initial model takes the following form:

$$
\begin{aligned}
C_i &\sim \text{Bernoulli}(logit(p_i)) \\
logit(p_i) &= \alpha + \beta_{FW} \times FW_i + \beta_{IW} \times IW_i \\
FW_i &\sim \text{Normal}(\mu, \sigma_{FW}) \\
\alpha &\sim \text{Normal}(0, 1) \\
\mu &\sim \text{Normal}(0, 1) \\
\beta_{FW} &\sim \text{Normal}(0, 1) \\
\beta_{IW} &\sim \text{Normal}(0, 1) \\
\sigma_{FW} &\sim \text{Exponential}(1)
\end{aligned}
\tag{3}
$$

Our analysis focuses on determining whether individuals successfully completed a diet program, denoted as C_i, where C_i equals 1 for success and 0 for dropout. To model this binary outcome, we employ a Bernoulli Distribution with a logit link function, a standard choice for logistic regression models. In our linear model, `logit_prob`, we utilize a log link function, incorporating an intercept, α and parameters β for all predictors.

One of our key predictors, `final_weight` (FW), contains missing values, requiring a nuanced approach. Expression (3) outlines the distribution for FW, accommodating both observed and unobserved missing values. When FW is observed, the equation functions as a likelihood, akin to a typical logistic regression. The model then learns the distributions of parameters, including μ and β_{FW}, consistent with the observed data. However, when FW is

missing, the equation serves as a prior, with parameters μ and β_{FW} estimated from the data, similar to varying effects in previous chapters. However, considering that these predictors represent the weight of the patient at different points in time, it seems unreasonable not to estimate the relationship between final and initial weight. Thus, our second model endeavors to address this by explicitly incorporating the association between these variables:

$$
\begin{aligned}
C_i &\sim \text{Bernoulli}(\text{logit}(p_i)) \\
\text{logit}(p_i) &= \alpha + \beta_{FW} \times FW_i + \beta_{IW} \times IW_i \\
(IW_i, FW_i) &\sim \text{MVNormal}((\mu_{IW}, \mu_{FW}), \mathbf{S}) \\
\alpha &\sim \text{Normal}(0, 1) \\
\mu_{IW}, \mu_{FW} &\sim \text{Normal}(0, 1) \\
\beta_{FW} &\sim \text{Normal}(0, 1) \\
\beta_{IW} &\sim \text{Normal}(0, 1) \\
\rho_{(IW,FW)} &\sim \text{lkj\_corr}(2) \\
\sigma_{(IW,FW)} &\sim \text{Exponential}(1)
\end{aligned}
\tag{4}
$$

In this second model, we incorporate a bivariate normal distribution that includes both final and initial weight in the imputation process, as opposed to solely relying on `final_weight`, the variable with missing values. The introduction of the **S** matrix represents another covariance matrix, which quantifies the correlation between final and initial weight using the observed cases. Specifically, the entries of the **S** matrix include a variance term (σ^2) for each variable's variance and a covariance term (ρ) representing the correlation between final and initial weight. This correlation information is then utilized to infer the missing final weight values. This approach ensures that our imputation process reflects the underlying association between these crucial predictors, enhancing the robustness and validity of our subsequent analyses.

For the entire diet models, we employed the standard configuration of 1000 samples distributed across 4 independent Markov chains. These chains were further allocated across 4 cores, facilitating concurrent execution across multiple processors rather than sequential processing. Consequently, the inferences derived automatically combined all non-warmup samples from each chain. However, for the phase-specific models, we opted for 2000 samples per chain. This decision was made because these models are less complex, resulting in reduced running time.

5.1.2 Age

Based on our literature review and descriptive statistics, age appears to be a significant factor associated with attrition in weight loss programs. Consequently, we incorporated this variable into our next model. Given the wide age range of participants, spanning from 16

to 84 years old, we categorized them into distinct groups: 16-25 years old, 25-50 years old, 50-75 years old, and 75-84 years old. For instances where age information was missing or unknown, we introduced a new category labeled as 'missing.' Subsequently, we represented age as dummy variables in our final model, resulting in the following specification:

$$
\begin{aligned}
C_i &\sim \text{Bernoulli}(\text{logit}(p_i)) \\
\text{logit}(p_i) &= \alpha + \beta_{\text{FW}} \times \text{FW}_i + \beta_{\text{IW}} \times \text{IW}_i + \\
&\quad \beta_{A^{25\_50}} \times A_i^{25-50} + \beta_{A^{50-75}} \times A_i^{50-75} + \\
&\quad \beta_{A^{75-84}} \times A_i^{75-84} + \beta_{A^{miss}} \times A_i^{miss} \\
(IW_i, FW_i) &\sim \text{MVNormal}((\mu_{\text{IW}}, \mu_{\text{FW}}), \mathbf{S}) \\
\alpha &\sim \text{Normal}(0, 1) \\
\mu_{\text{IW}}, \mu_{\text{FW}} &\sim \text{Normal}(0, 1) \\
\beta_{\text{FW}}, \beta_{\text{IW}}, \beta_{A^{25\_50}}, \beta_{A^{50-75}}, \beta_{A^{75-84}}, \beta_{A^{miss}} &\sim \text{Normal}(0, 1) \\
\rho_{(IW,FW)} &\sim \text{lkj\_corr}(2) \\
\sigma_{(IW,FW)} &\sim \text{Exponential}(1)
\end{aligned}
\tag{5}
$$

In the updated model, the only modification from the previous one is the inclusion of additional variables: A25-50 (equal to 1 if an individual's age falls between 25 and 50 years old), A50-75 (indicating individuals aged between 50 and 75 years old), A75-84 (representing individuals aged between 75 and 84 years old), and a category denoting missing age data. To circumvent the dummy variable trap, we have removed the dummy variable for the 16-25 age group, which encompasses individuals between 16 and 25 years old.

5.1.3 Final Model

In our final model, we incorporated dummy variables for the phase of the diet program. Our descriptive statistics revealed clear variations in attrition rates across different phases, indicating that phase plays a significant role in determining attrition. Additionally, we opted to include other variables that, while not highly significant in the descriptive statistics, showed some degree of influence. These variables encompassed gender, the number of diet program repetitions, and the specific program purchased by the individual (e.g., loose consultation, 4-week pack, 8-week pack, second phase pack, twelve weeks pack, and balance pack).

$$
\begin{aligned}
C_i &\sim \text{Bernoulli}(\text{logit}(p_i)) \\
\text{logit}(p_i) &= \alpha + \beta_{\text{FW}} \times \text{FW}_i + \beta_{\text{IW}} \times \text{IW}_i + \\
&\quad \beta_{\text{Age}} \times \text{Age}_i + \beta_{\text{Phase}} \times \text{Phase}_i + \\
&\quad \beta_{\text{Program}} \times \text{Program}_i + \beta_{\text{Repeated}} \times \text{Repeated}_i \\
(IW_i, FW_i) &\sim \text{MVNormal}((\mu_{\text{IW}}, \mu_{\text{FW}}), \mathbf{S}) \\
\alpha &\sim \text{Normal}(0, 1) \\
\mu_{\text{IW}}, \mu_{\text{FW}} &\sim \text{Normal}(0, 1) \\
\beta_{\text{FW}}, \beta_{\text{IW}}, \beta_{\text{Age}}, \beta_{\text{Phase}}, \beta_{\text{Program}}, \beta_{\text{Repeated}} &\sim \text{Normal}(0, 1) \\
\rho_{(IW,FW)} &\sim \text{lkj\_corr}(2) \\
\sigma_{(IW,FW)} &\sim \text{Exponential}(1)
\end{aligned}
\tag{6}
$$

In model (6) the variables Age, Phase, Program and repeated represent the dummy variables for `age_groups`, `phase`, `program`, and `repeated`, respectively. Furthermore, we acknowledge the potential impact of additional variables such as `brand_adhesion_type`, which, as mentioned in the descriptive statistics, also demonstrate slight variations across categories. Additionally, factors like `profession` and `location` may exert influence. These factors could shed light on variations in disposable income and lifestyle preferences, as individuals from different professions and regions may allocate resources differently. For instance, individuals residing in affluent areas like Lisbon and Porto might exhibit higher spending capacities for such services. Similarly, certain professions, particularly those linked to fitness or aesthetics, might display heightened interest in weight loss programs, for example. Moreover, exploring marketing-related variables presents an intriguing avenue for the company. Understanding how clients become aware of the diet program, whether through clinic visits, online platforms, social media, influencers, radio, or television, could provide valuable insights into their engagement and commitment levels.

While these variables offer rich potential for exploration, practical constraints limited our ability to analyze them comprehensively. Therefore, we focused on variables that demonstrated notable impacts in our descriptive statistics, underscoring the significance of this analytical approach.

5.2 Per Phase

Given the substantial variation in attrition observed across different phases of our dietary intervention, we have chosen to delve deeper into understanding completion rates within each phase. Leveraging the model we have just discussed, we aim to gain a comprehensive un-

derstanding of the dietary program while examining if the observed patterns hold true across individual phases and whether they carry similar significance.

Previously, we defined attrition and its implications. To model completion within each phase, we partitioned our dataset into three distinct tables based on phase designation. Each table encompasses consultations specific to its corresponding phase, utilizing the initial weights recorded at the onset of each respective phase. This approach essentially treats each phase as an independent process.

While this strategy enables focused analysis within each phase, it comes with potential drawbacks. By exclusively considering patients who progressed beyond the initial phase, we inadvertently exclude individuals who did not transition to subsequent phases. Moreover, we overlook pertinent historical data for patients within each phase. For instance, when analyzing observations solely from phase 2, we disregard all information related to phase 1 for the same patient. Therefore it is crucial to emphasize that we treat these analyses as almost separate programs. Analysis of phase 2, for instance, only involves patients actively participating in that phase, thereby serving as a complement to our overarching analysis. The latter, which encompasses the entire dataset, ensures that no information or patient is excluded.

Nonetheless, exploring specific patterns within each phase remains valuable. Understanding if distinct trends emerge within individual phases can offer valuable insights into program dynamics and participant behavior.

6 Results

In this section, we will outline the results obtained from our models. Initially, we will present the findings pertaining to the complete weight loss program, followed by an analysis of each individual phase.

6.1 Complete Diet

6.1.1 Weight

As discussed, our initial model focuses on leveraging weight, specifically initial weight and final weight, as predictive factors. However, integrating final weight poses an extra challenge due to missing data. Figure 5 illustrates the inferred distribution between the two predictors, `final_weight` and `initial_weight`, alongside our imputed values. It's evident that the imputation process did not yield the expected results.

Firstly, it's worth noting that for individuals with the same initial weight, the imputed final weights tend to be significantly higher than the final weights observed. This discrepancy could be rationalized by considering that those who drop out of the study may exhibit lower motivation, reduced availability for consultations, or even constrained financial resources, resulting in less involvement in the diet and therefore, greater weight gain over time. However, despite a discernible positive correlation between the two predictors, the imputed values do not adequately capture this relationship. This incongruity stems from the underlying assumption of the imputation model, which posits no inherent relationship between the two predictors.

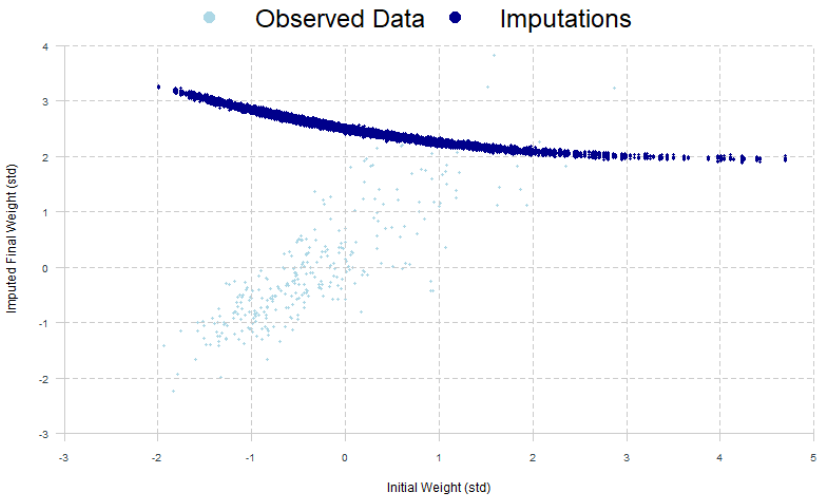


Figure 5: Infered Distribution of Predictors for Weight Model

Note: The graph displays the inferred distribution of final weight (vertical) and initial weight (horizontal), with imputed values shown by dark blue points.

However, both graphically and logically, we can observe a positive correlation between these predictors. This suggests that higher initial weights tend to correlate with higher final weights, and vice versa. Such a correlation could imply that customers who begin with higher weights are more likely to retain higher weights, potentially impacting completion rates.

The robustness of our prediction results, as showcased in Table 1, is underscored by the convergence of our Markov chain, as evidenced by Gelman-Rubin statistics (Rhat)[2] approaching 1 and high number of effective sample (ess_bulk)[3]. With 1000 samples utilized in our Markov

[2]The Rhat measures the convergence of multiple chains in a MCMC sampler. A value close to 1 indicates convergence to the stationary distribution. Acceptable values typically fall in the neighborhood of 1.1 or 1.2. A value of 1 indicates perfect convergence, where all chains are essentially indistinguishable from each other and have effectively sampled from the stationary distribution .

[3]The ess_bulk is an estimate of the number of independent samples from the posterior distribution. Markov chains often exhibit autocorrelation, meaning sequential samples are not entirely independent, thereby reducing the effective number of samples. Acceptable ess_bulk values should not be much lower than the actual number of iterations (minus warmup) of our chains.(McElreath, 2023)

chain for the entirety of the diet, this provides a solid foundation for confidence in our model estimates.

Table 1: Posterior Distribution for Parameters of Interest for Weight Model

	Mean	**Standard Deviation**	**5.5%**	**94.5%**	**Rhat**	**ess_bulk**
ν	-1.94	0.02	-1.97	-1.91	1.02	131.45
α	3.44	0.10	3.29	3.61	1.00	729.16
β_{Iw}	4.69	0.11	-4.88	-4.52	1.00	886.35
β_{Fw}	-6.39	0.13	6.18	6.60	1.00	942.78
σ_{Fw}	1.63	0.01	1.60	1.65	1.02	128.64

These results reveal two primary trends: initial weight and final weight have opposing effects on completion rates, which is consistent with the nature of a diet program. Specifically, individuals with higher initial weights are more likely to complete the diet, likely due to greater motivation and potentially easier weight loss. Conversely, higher final weights correspond to lower completion rates, aligning with the primary goal of weight loss in the diet program.

These findings not only align with our graphical representations but also reflect the inherent dynamics of weight loss endeavors. Overall, our analysis underscores the validity of our model and offers valuable insights into the factors influencing completion rates within our diet program.

In our second model, as previously mentioned, we incorporated the positive relationship between β_{Fw} and β_{Iw}. The results of the inferred distribution between these two predictors are shown in Figure 6. While they still do not completely capture the observed data, they show improvement compared to the first model. This progress is encouraging as a starting point, indicating potential for further refinement.

In the figure, we can see that a trend emerges: individuals with the same initial weight tend to have a lower final weight if they drop out of the program. This observation prompts an intriguing inquiry: what might be the underlying reasons for this pattern? Several potential explanations exist. Weight loss is known to be non-linear, with individuals experiencing fluctuations throughout their journey. Those who complete may have initially lost weight but subsequently regained it before leaving the program, resulting in a higher final weight compared to those who completed the program. Additionally, we must consider the possibility of sample bias or confounding variables influencing the observed relationship between dropout status and final weight. Factors such as age, gender, socioeconomic status, or comorbidities could interact with completion status and impact final weight differently, as we will delve into further in our investigation.

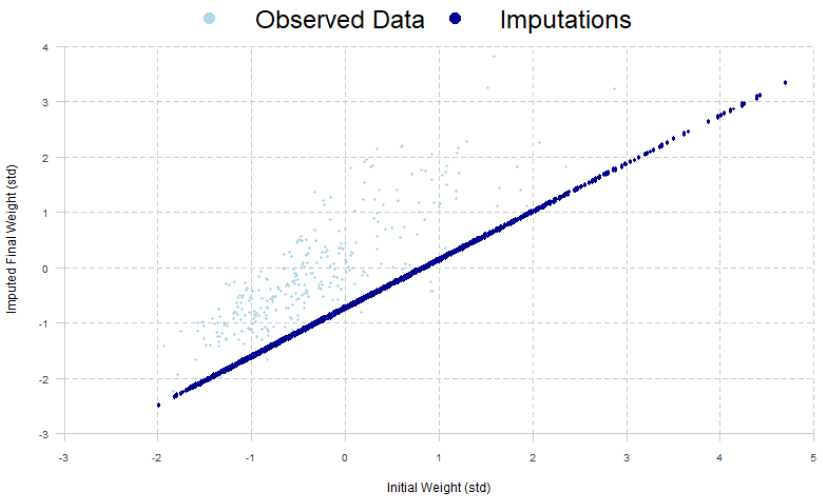


Figure 6: Infered Distribution of Predictors Accounting For Correlation

Note: The graph displays the inferred distribution of final weight (vertical) and initial weight (horizontal), with imputed values shown by dark blue points.

Table 2 presents the metrics of our Bayesian inference model.

Table 2: Posterior Distribution for Parameters of Interest for Weight

	Mean	**Standard Deviation**	**5.5%**	**94.5%**	**Rhat**	**ess_bulk**
μ_{Iw}	-0.78	0.26	-0.96	-0.19	1.51	7.55
μ_{Fw}	-0.00	0.01	-0.01	0.01	1.01	3669.73
α	0.53	0.79	-1.16	1.12	1.50	7.59
β_{Iw}	4.07	0.99	-4.75	-1.77	1.50	7.54
β_{Fw}	-4.31	1.19	1.60	5.14	1.50	7.51
σ	1.02	1.08	0.05	2.98	1.00	2494.87
$\rho_{\text{FwM}[1,2]}$	0.72	0.06	0.67	0.85	1.51	7.52
$\sigma_{\text{FwM}[1]}$	1.00	0.00	0.99	1.01	1.00	2846.75
$\sigma_{\text{FwM}[2]}$	1.17	0.05	1.07	1.21	1.51	7.41

Based on the Rhat values exceeding 1 and the relatively low ess_bulk values, it appears that this model may not be very reliable. The Rhat values indicate a lack of convergence, suggesting that additional samples are necessary to obtain reliable inference. Additionally, the low ess_bulk values indicate that the sampling process is not very efficient, resulting in limited independent information obtained from the chains. Having a sufficiently large effective sample size is crucial for ensuring the reliability of parameter estimation and credible intervals in Bayesian inference. Therefore, further investigation and potentially increasing the number of samples or improving the sampling algorithm may be necessary to enhance the reliability of the model results.

In the subsequent models, we introduce additional variables not only to enhance churn

prediction but also to indirectly aid the imputation of Fw. These variables offer supplementary information that can contribute to predicting missing values of Fw. While they may not directly represent Fw, they could exhibit correlations or contain pertinent details that assist in imputing missing values. By incorporating this broader set of predictors, we aim to improve the accuracy of our imputation model and enhance our understanding of the factors influencing Fw.

6.1.2 Age

As shown in Figure 7, incorporating age into the model does not appear to bring the imputed mean values closer to the actual observed values for initial weight (Iw). Notably, the imputed final weight for individuals who discontinued the program appears to be lower when compared to the observed final weight for individuals with the same initial weight.

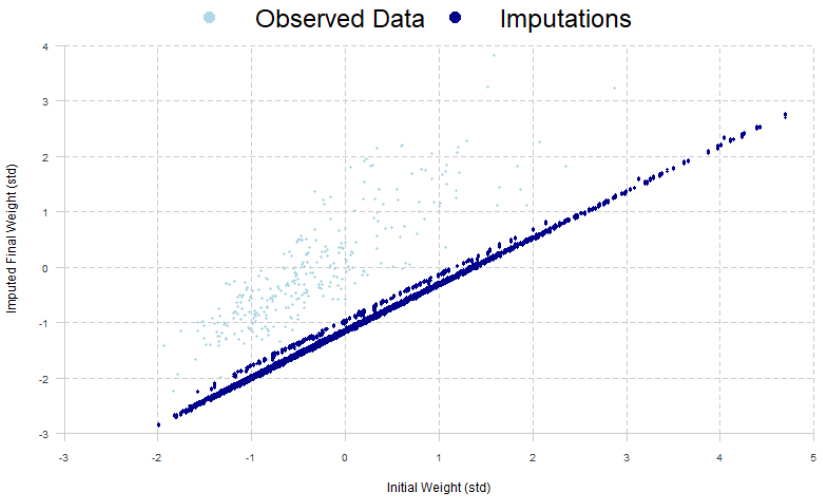


Figure 7: Infered Distribution of Predictors Accounting For Age Model

Note: The graph displays the inferred distribution of final weight (vertical) and initial weight (horizontal), with imputed values shown by dark blue points.

Table 3 presents the outcomes derived from our third model, aimed at predicting diet completion using `initial_weight`, `final_weight` and `age` as predictors and provides insights into completion tendencies based on the estimated coefficients:

1. Initial Weight (β_{Iw}): The estimated coefficient for initial weight indicates a positive association with completion. This suggests that higher initial weights are linked to higher completion probabilities, implying that customers with higher initial weights, possibly indicating greater investment or commitment, are more likely to complete the diet.

2. Final Weight (β_{Fw}): The estimated negative coefficient for final weight suggests that lower final weights are associated with higher completion rates. This observation aligns with the common goal of diet programs, which is typically centered around weight loss.

Table 3: Posterior Distribution for Parameters of Interest for Age Mode

Parameter	Mean	Standard Deviation	5.5%	94.5%	Rhat	ess
μ_{Iw}	-0.84	0.02	-0.87	-0.80	1.05	53.64
μ_{Fw}	0.00	0.01	-0.01	0.01	1.00	3886.83
α	-2.28	0.19	-2.60	-1.98	1.01	791.24
$\beta_{\text{Age\_missing}}$	-0.01	1.03	-1.68	1.61	1.00	3531.13
$\beta_{\text{Age\_75}}$	0.87	0.28	0.44	1.34	1.00	1107.06
$\beta_{\text{Age\_50}}$	3.10	0.20	2.80	3.41	1.01	852.84
$\beta_{\text{Age\_25}}$	2.81	0.20	2.49	3.13	1.01	775.90
β_{Iw}	-3.96	0.09	-4.10	-3.82	1.03	136.67
β_{Fw}	4.32	0.11	4.14	4.51	1.04	88.34
σ	0.99	0.90	0.08	2.69	1.00	2046.82
$\rho_{\text{FwM[1,2]}}$	0.68	0.01	0.67	0.69	1.04	64.08
$\Sigma_{\text{FwM[1]}}$	1.00	0.00	0.99	1.01	1.00	3045.24
$\Sigma_{\text{FwM[2]}}$	1.17	0.01	1.16	1.19	1.03	90.94

Individuals who experience greater weight loss are likely more satisfied with the program, leading to higher motivation and increased likelihood of completing the diet.

3. Age Group Effects ($\beta_{\text{Age\_25}}$, $\beta_{\text{Age\_50}}$, $\beta_{\text{Age\_75}}$, $\beta_{\text{Age\_missing}}$): We anticipated that older individuals would be more likely to complete the diet program, in line with existing literature and our own descriptive statistics. Consistent with our expectations, individuals in the age groups of 25 to 50 years old, 50 to 75 years old, and 75 to 84 years old exhibit a higher likelihood of completing the diet compared to those aged 16 to 25, as evidenced in the table. However, it is noteworthy that the magnitude of this effect appears to be less pronounced for individuals aged more than 75. The strongest effect is observed among individuals in their 50s to 75s, followed by those in the 25 to 50 age group, with the effect tapering off for individuals aged more than 75. This suggests that while age does influence completion rates, its impact may vary across different age groups.

4. Correlation Between Initial and Final Weights ($\rho_{\text{FwM[1,2]}}$): The positive correlation coefficient between initial and final weights confirms their strong positive association (approximately 0.68). This indicates that higher initial weights are associated with higher final weights, and vice versa, as we previously concluded.

5. Regression Coefficients (β_{Iw}, β_{Fw}): These coefficients demonstrate the impact of initial and final weights on churn when controlling for other factors. Positive coefficients for initial weight and negative coefficients for final weight suggest that higher initial weights and lower final weights are correlated with higher completion probabilities.

Once again, the Gelman-Rubin statistic (Rhat) and effective sample size (ess) present robust results, indicating that we may trust these predictions.

6.1.3 Final Model

As depicted in Figure 8, the imputations generated by our final model consistently yield lower values compared to the observed values. While we have previously discussed potential reasons for this phenomenon, it's crucial to acknowledge that our dataset, now encompassing the entire program, is highly imbalanced with skewed class proportions. This skewness could potentially introduce interference with our imputation process.

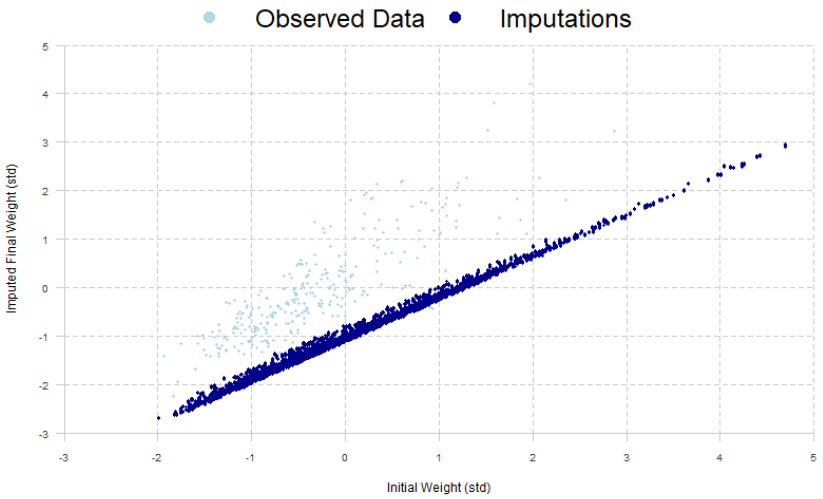


Figure 8: Infered Distribution of Predictors Accounting For Final Diet Model

Note: The graph displays the inferred distribution of final weight (vertical) and initial weight (horizontal), with imputed values shown by dark blue points.

Table 4 presents the metrics results for the last model. We will carefully examine some of the conclusions we can draw from these results:

- Initial and Final weight maintain a very intense effect on completion similar with previous results.
- Brand Adhesion Type (β_{second}, $\beta_{\text{weeks\_4}}$, $\beta_{\text{weeks\_8}}$, $\beta_{\text{weeks\_12}}$, $\beta_{\text{weeks\_6}}$, β_{balance} β_{first}): The omitted dummy variable is the loose consultation. Our analysis suggests that all brand adhesion plans have lower completion probabilities compared to loose consultations, as indicated by their negative estimated coefficients.
- Gender (β_{male}): The estimated coefficient for gender is close to zero, indicating that gender might not have a strong association with completion once other factors are considered. However, further investigation may be warranted to fully understand the role of gender in completion rates.
- Diet Repetition (β_{repeated}): The estimated coefficient for repeating the diet is negative, indicating that customers with that repeated the diet at some point in time, are more

Table 4: Posterior Distribution for Parameters of Interest for Final Model

	Mean	**Standard Deviation**	**5.5%**	**94.5%**	**Rhat**	**ess_bulk**
μ_{Iw}	-0.75	0.02	-0.78	-0.70	1.10	32.78
μ_{Fw}	-0.00	0.00	-0.01	0.01	1.00	2604.06
α	-2.01	0.18	-2.30	-1.72	1.01	529.37
β_{Iw}	-3.75	0.08	-3.88	-3.62	1.04	100.17
β_{Fw}	4.03	0.11	3.86	4.20	1.06	62.62
$\beta_{\text{phase\_3}}$	-0.00	0.99	-1.55	1.63	1.00	4030.23
$\beta_{\text{phase\_2}}$	-0.01	0.99	-1.63	1.57	1.00	4329.32
β_{male}	0.02	1.02	-1.63	1.67	1.00	2868.97
$\beta_{\text{weeks\_6}}$	-1.38	0.78	-2.62	-0.15	1.00	2288.86
β_{balance}	-0.02	1.03	-1.67	1.61	1.01	4581.34
$\beta_{\text{weeks\_12}}$	-1.14	0.30	-1.61	-0.66	1.00	2099.03
β_{second}	-0.45	0.93	-1.93	1.04	1.00	2925.04
$\beta_{\text{weeks\_8}}$	-4.45	0.59	-5.45	-3.55	1.00	1998.83
$\beta_{\text{weeks\_4}}$	-1.01	0.09	-1.15	-0.87	1.00	1129.86
β_{first}	-0.49	0.18	-0.78	-0.20	1.00	2593.81
β_{repeated}	-0.22	0.01	-0.24	-0.20	1.01	378.11
$\beta_{\text{Age\_missing}}$	0.02	0.98	-1.51	1.57	1.00	2990.59
$\beta_{\text{Age\_75}}$	1.03	0.26	0.62	1.44	1.01	810.86
$\beta_{\text{Age\_50}}$	3.18	0.18	2.90	3.46	1.00	801.29
$\beta_{\text{Age\_25}}$	2.83	0.18	2.55	3.11	1.00	720.32
$\beta_{\text{Phase\_3}}$	-0.02	1.03	-1.65	1.59	1.00	3848.63
$\beta_{\text{Phase\_2}}$	0.02	0.95	-1.50	1.52	1.00	3504.62
σ	0.99	0.96	0.05	2.79	1.00	2600.00
$\rho_{\text{FwM[1,2]}}$	0.71	0.01	0.69	0.72	1.10	35.79
$\sigma_{\text{FwM[1]}}$	1.00	0.00	0.99	1.01	1.00	2735.13
$\sigma_{\text{FwM[2]}}$	1.15	0.01	1.13	1.16	1.03	107.41

likely to churn compared to a reference category, which in this case is people that do not repeat the diet, the effect is not very strong though.

- Age Group Effects ($\beta_{\text{Age\_25}}$, $\beta_{\text{Age\_50}}$, $\beta_{\text{Age\_75}}$, $\beta_{\text{Age\_missing}}$): For age, we still find the same pattern as in the previous model.

- Phase ($\beta_{\text{Phase\_2}}$, $\beta_{\text{Phase\_3}}$): The estimated coefficient for both phases is close to zero, indicating that these phases might not have a strong association with completion once other factors are considered. However, further investigation may be warranted to fully understand the role of phase in completion rates.

Once again, our results demonstrate robustness, as indicated by the Rhat and ess_bulk metrics.

Bayesian methods also provide robust estimates for uncertainty. The 5.5% and 94.5% quantiles delineate the boundaries of a percentile interval, representing an 89% compatibility interval. If the credible interval for a parameter excludes zero, it implies that the parameter esti-

mate significantly differs from zero with an 89% level of confidence. Our findings in Table 4 demonstrate that for most parameters, particularly those of interest such as age and weight, our estimates diverge from zero with an 89% level of confidence.

To deepen our analysis of chain health, we will also include a trace plot displayed in Figure 9. Looking at the trace plot of each parameter is often the best way to diagnose common problems with our chains.

A trace plot is essentially a time series plot where the y-values represent the samples generated by the chain. It visually displays the sequential order of these samples, connected by a line (McElreath, 2023). These plots serve as quick tools to assess whether the chain has reached its target distribution. Once the chain has ”burned in”, indicating that it has reached a stable state and is effectively sampling from the target distribution, the trace plots should show rapid oscillation around a central value. This suggests that the chain is exploring the target distribution thoroughly, providing reliable estimates of the parameters of interest (McElreath, 2023).

According to Figure 9, our Markov chains appear to be in good condition, exhibiting both stationarity and effective mixing. There are no evident anomalies observed. However, we do notice a slight serial correlation between successive draws, and the chains seem to explore the sample space thoroughly. It is worth noting that certain age and weight variables display a noticeable discrepancy between the initial and subsequent parts of the sample. This suggests that the initial distribution and subsequent distributions of the chain were initially dissimilar to the target distribution, but gradually converged over time.

This discrepancy could pose a concern as it seems that a substantial portion of the sample is drawn from distributions that significantly differ from the target distribution. Despite our efforts to ensure representativeness in our random sampling of data, such variations can occur.

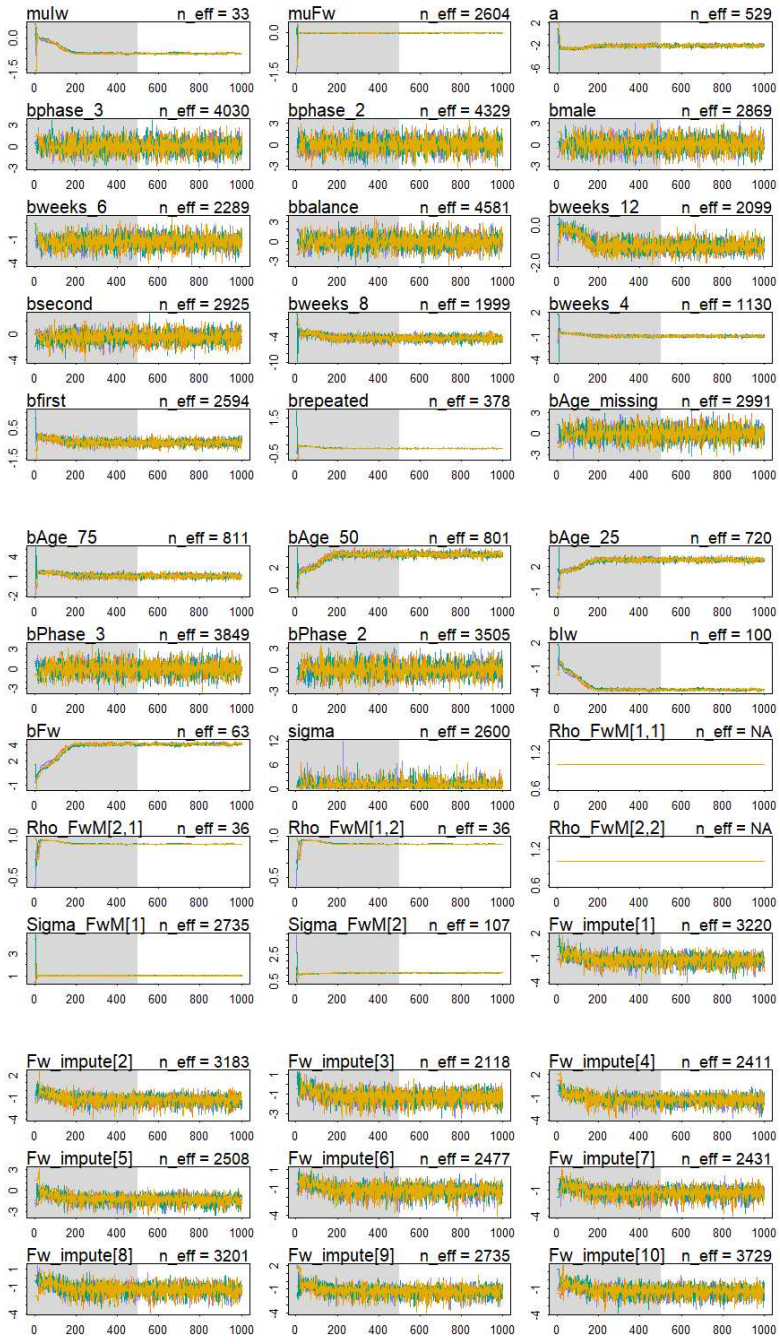


Figure 9: Trace Plot of the Markov chains for Final Model

Note: The figure displays Trace plot of the Markov chain from the final model of the whole diet. The gray region is warmup, during which the Markov chain was adapting to improve sampling efficiency. The white region contains the samples used for inference.

6.2 Per Phase

As previously indicated, our analysis will now focus on individual phases. We will present the imputation results, attrition prediction outcomes, and assess the convergence of the Markov Chains.

6.2.1 Phase 1

We begin our analysis with phase 1. Figure 10 illustrates a robust positive correlation between final and initial weight, which our model adeptly captures.

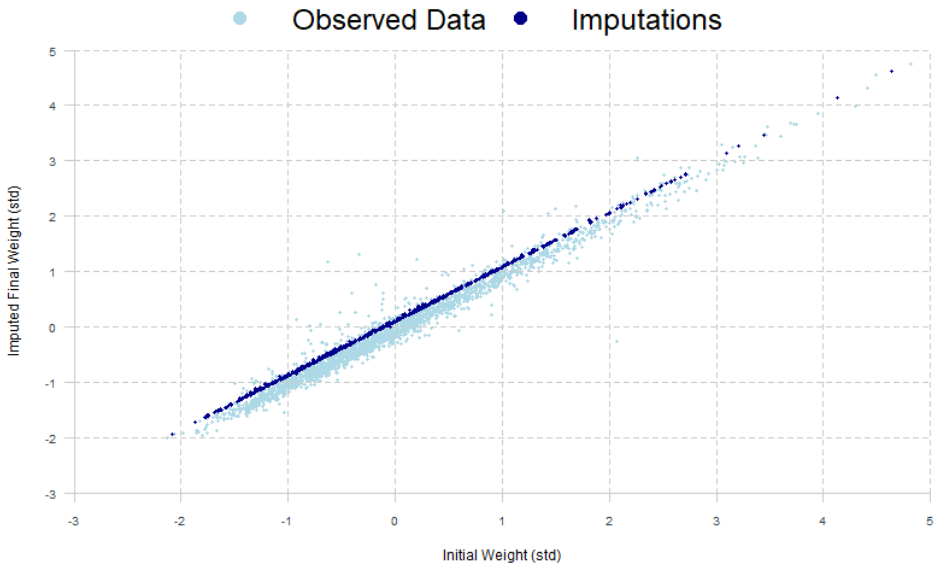


Figure 10: Infered Distribution of Weight Predictors for Phase 1

Note: The graph displays the inferred distribution of final weight (vertical) and initial weight (horizontal), with imputed mean values shown by dark blue points.

Additionally, our model suggests that individuals who churned in Phase 1 tend to exhibit higher final weights for the same initial weight. This conclusion appears reasonable considering that Phase 1 is the initial and shortest phase, marked by stringent dietary requirements and reductions in food intake. It serves as the first significant impact on participants' weight loss journey. Therefore, if patients fail to perceive results after making dietary sacrifices, it is reasonable to expect attrition.

The predictive factors influencing attrition are summarized in Table 3 and indicate that the results from phase 1 closely resemble those of the entire diet program.

However, there are notable new trends:

- Brand Adhesion Type: The coefficients for brand adhesion type are now positive, suggesting that individuals with these plans are more likely to complete the program com-

Table 5: Posterior Distribution for Parameters of Interest for Phase 1

	Mean	**Standard Deviation**	**5.5%**	**94.5%**	**Rhat**	**ess_bulk**
μ_{Iw}	0.01	0.01	-0.01	0.02	1.00	4049.85
μ_{Fw}	-0.00	0.01	-0.02	0.02	1.00	6542.98
α	0.45	0.12	0.27	0.64	1.00	3774.93
β_{male}	0.02	1.01	-1.61	1.67	1.00	9073.38
$\beta_{\text{weeks\_6}}$	-1.81	0.72	-2.97	-0.68	1.00	9864.65
β_{balance}	0.02	0.99	-1.58	1.59	1.00	8686.57
$\beta_{\text{weeks\_12}}$	0.93	0.57	0.07	1.86	1.00	9750.06
β_{second}	-1.86	0.70	-2.99	-0.77	1.00	9473.13
$\beta_{\text{weeks\_8}}$	0.06	0.31	-0.43	0.56	1.00	7515.81
$\beta_{\text{weeks\_4}}$	0.15	0.08	0.02	0.28	1.00	7410.67
β_{first}	0.37	0.23	0.02	0.72	1.00	9461.85
β_{repeated}	0.28	0.02	0.25	0.31	1.00	2792.72
$\beta_{\text{Age\_missing}}$	0.01	1.03	-1.67	1.64	1.00	9926.86
$\beta_{\text{Age\_75}}$	0.63	0.32	0.12	1.17	1.00	7041.08
$\beta_{\text{Age\_50}}$	0.23	0.12	0.04	0.42	1.00	4093.33
$\beta_{\text{Age\_25}}$	0.00	0.12	-0.18	0.19	1.00	3821.42
β_{Iw}	4.32	0.33	3.79	4.84	1.00	848.57
β_{Fw}	-4.27	0.33	-4.80	-3.72	1.00	842.88
σ	1.00	1.02	0.05	2.98	1.00	5693.40
$\rho_{\text{FwM[2,1]}}$	0.98	0.00	0.98	0.98	1.00	1883.06
$\sigma_{\text{FwM[1]}}$	1.00	0.01	0.99	1.01	1.00	6290.84
$\sigma_{\text{FwM[2]}}$	0.99	0.01	0.98	1.00	1.00	6517.31

pared to those with loose consultations. This finding aligns with the phase's structure, as it involves fewer consultations. Therefore, individuals who purchase a plan are more inclined to attend at least the initial consultations and complete phase 3. Notably, the coefficient with the highest value is for β_{first}, representing clients who purchased the entire first phase upfront. It's logical that these individuals are more likely to complete the first phase.

- Age: A noticeable trend emerges regarding age, during Phase 1, the observed pattern aligns well with our expectations based on the literature review. Specifically, older individuals demonstrate a higher likelihood of completion compared to patients aged 16-25. Furthermore, this likelihood appears to increase progressively with each older age group.

The traceplots for analyzing the validity of our Markov chains for the principal variables in Phase 1 are provided in the appendix. Overall, there are no apparent anomalies for most variables, which aligns with the good results obtained from the effective number of samples and Rhat. We observe mild serial correlation between successive draws, and the chains appear to explore the sample space thoroughly.

However, for coefficients such as β_{IW} and β_{FW}, the initial part of the sample exhibits noticeable differences from the remaining part. This discrepancy suggests that the initial distri-

bution and the subsequent distributions of the chain may have been markedly different from the target distribution. Nonetheless, the chain gradually converged to the target distribution over time. While this discrepancy occurs in only a small portion of our data, it does not appear to pose a significant issue.

6.2.2 Phase 2

Moving on to phase 2, a notable observation contrasts with the outcomes of phase 1 as we can see in Figure 11.

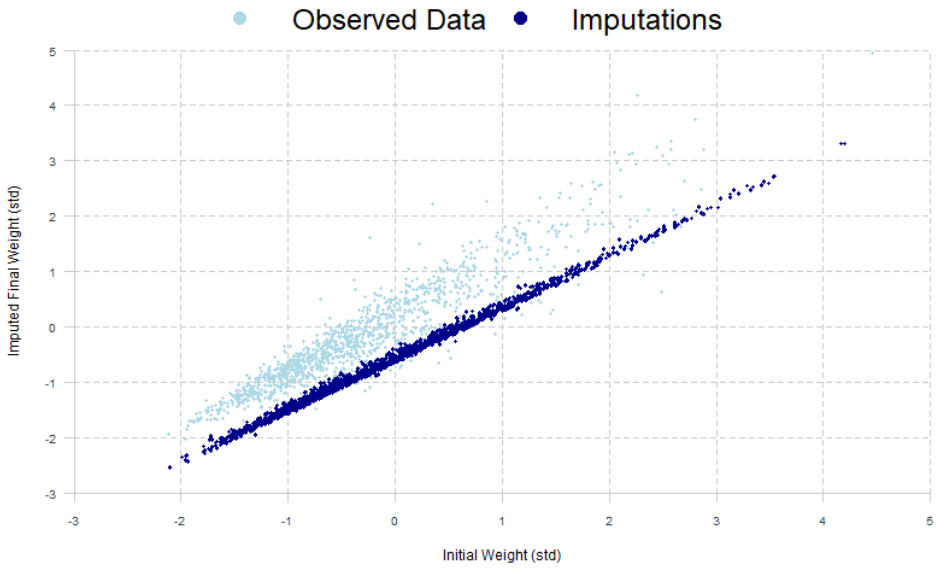


Figure 11: Infered Distribution of Weight Predictors for Phase 2

Note: The graph displays the inferred distribution of final weight (vertical) and initial weight (horizontal), with imputed mean values shown by dark blue points.

Specifically, the imputed final weights appear to be lower than the observed weights for individuals with the same initial weight. In other words, individuals who do not complete phase 2 seem to exhibit lower weights, as depicted in Figure 11. One plausible explanation for this phenomenon may be attributed to the extended duration of phase 2. It is conceivable that individuals who achieve significant weight loss and reach their goals may opt to discontinue the program prematurely.

Table 5 displays the results for the posterior distribution for parameters of interest for Phase 2 and we can see that in this phase, most trends remain consistent, but a notable deviation is observed in the coefficients related to brand adhesion plans.

Our analysis indicates that both the 6-week plan and the second phase plan show lower completion probabilities compared to loose consultations, as evidenced by their negative esti-

Table 6: Posterior Distribution for Parameters of Interest for Phase 2

	Mean	**Standard Deviation**	**5.5%**	**94.5%**	**Rhat**	**ess_bulk**
μ_{Iw}	0.01	0.01	-0.01	0.02	1.00	4049.85
μ_{Fw}	-0.00	0.01	-0.02	0.02	1.00	6542.98
α	0.45	0.12	0.27	0.64	1.00	3774.93
β_{male}	0.02	1.01	-1.61	1.67	1.00	9073.38
$\beta_{\text{weeks\_6}}$	-1.81	0.72	-2.97	-0.68	1.00	9864.65
β_{balance}	0.02	0.99	-1.58	1.59	1.00	8686.57
$\beta_{\text{weeks\_12}}$	0.93	0.57	0.07	1.86	1.00	9750.06
β_{second}	-1.86	0.70	-2.99	-0.77	1.00	9473.13
$\beta_{\text{weeks\_8}}$	0.06	0.31	-0.43	0.56	1.00	7515.81
$\beta_{\text{weeks\_4}}$	0.15	0.08	0.02	0.28	1.00	7410.67
β_{first}	0.37	0.23	0.02	0.72	1.00	9461.85
β_{repeated}	0.28	0.02	0.25	0.31	1.00	2792.72
$\beta_{\text{Age\_missing}}$	0.01	1.03	-1.67	1.64	1.00	9926.86
$\beta_{\text{Age\_75}}$	0.63	0.32	0.12	1.17	1.00	7041.08
$\beta_{\text{Age\_50}}$	0.23	0.12	0.04	0.42	1.00	4093.33
$\beta_{\text{Age\_25}}$	0.00	0.12	-0.18	0.19	1.00	3821.42
β_{Iw}	4.32	0.33	3.79	4.84	1.00	848.57
β_{Fw}	-4.27	0.33	-4.80	-3.72	1.00	842.88
σ	1.00	1.02	0.05	2.98	1.00	5693.40
$\rho_{\text{FwM[2,1]}}$	0.98	0.00	0.98	0.98	1.00	1883.06
$\sigma_{\text{FwM[1]}}$	1.00	0.01	0.99	1.01	1.00	6290.84
$\sigma_{\text{FwM[2]}}$	0.99	0.01	0.98	1.00	1.00	6517.31

mated coefficients. This finding may seem unexpected and warrants special attention from the company. For Phase 2, which is designed for long-term sustainability, one would anticipate a high completion rate among individuals who purchase all consultations for this phase, as it represents a significant incentive. Similarly, for the 6-week plan, which offers a substantial duration of support, one would expect a higher completion rate.

One plausible explanation for the observed discrepancy could be that these plans are relatively new, and individuals who have purchased them may not have had sufficient time to complete the diet regimen. As a result, our sample may not be sufficiently balanced or representative for these specific plans, leading to the unexpected findings. This aspect should be carefully considered and investigated further by the company.

6.2.3 Phase 3

Lastly, we arrive at Phase 3, and the corresponding results for imputation are displayed in Figure 12.

The observed trend in Phase 2 persists in Phase 3, wherein clients who do not complete the phase appear to have a final weight lower than those who do complete it, despite having the

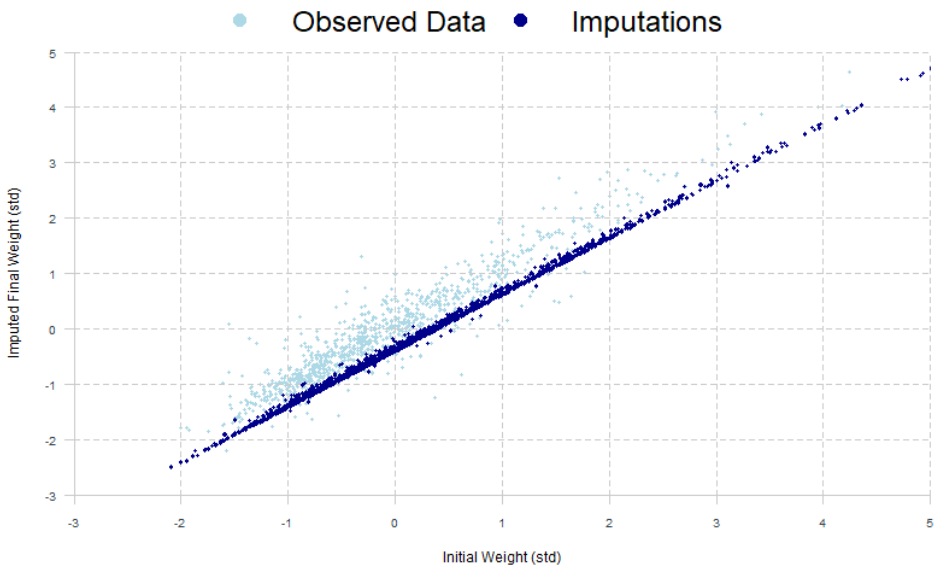


Figure 12: Infered Distribution of Weight Predictors for Phase 3

Note: The graph displays the inferred distribution of final weight (vertical) and initial weight (horizontal), with imputed mean values shown by dark blue points.

same initial weight. A potential explanation for this pattern is akin to that of Phase 2. Phase 3 primarily serves as a maintenance phase, during which patients are expected to adopt and practice healthy habits to sustain their weight in the long term. In this phase, patients are not necessarily anticipated to continue losing weight, rather, the focus is on weight maintenance. Given the extended duration of Phase 3 (lasting 8 months), it is plausible that clients who drop out are those who feel confident enough to discontinue the diet and begin the independent process earlier, thereby resulting in lower final weights.

Regarding the inference results, Table 6 highlights significant changes, particularly in the coefficients related to brand adhesion type. Notably, the 6-week plan, second phase plan, 8-week plan, 4-week plan, and first phase plan now exhibit negative coefficients, suggesting lower probabilities of completion compared to loose consultations.

It's important to emphasize that these packs are typically marketed during Phase 2, which is often the longest phase where clients are still actively expected to lose weight. The negative coefficients for these plans may indicate unexpected challenges or barriers faced by individuals who opt for these packages. In contrast, Phase 3 serves as a maintenance phase. While it requires 8 consultations for clients to officially complete, it's plausible to assume that individuals may prefer loose consultations during this phase, allowing them to discontinue once they feel confident in maintaining their weight loss progress.

Table 7: Posterior Distribution for Parameters of Interest for Phase 1

	Mean	**Standard Deviation**	**5.5%**	**94.5%**	**Rhat**	**ess_bulk**
μ_{Iw}	-0.08	0.01	-0.09	-0.06	1.01	1265.77
μ_{Fw}	-0.00	0.01	-0.01	0.01	1.00	4635.88
α	0.27	0.13	0.07	0.47	1.01	1389.30
β_{male}	0.04	1.01	-1.56	1.64	1.00	6381.43
$\beta_{\text{weeks\_6}}$	-1.62	0.41	-2.28	-1.00	1.00	4818.99
β_{balance}	0.01	0.99	-1.59	1.62	1.00	6355.65
$\beta_{\text{weeks\_12}}$	0.10	0.32	-0.41	0.62	1.00	5830.54
β_{second}	-0.31	0.32	-0.82	0.22	1.00	5331.37
$\beta_{\text{weeks\_8}}$	-4.09	0.55	-4.99	-3.22	1.00	5225.66
$\beta_{\text{weeks\_4}}$	-1.31	0.10	-1.47	-1.15	1.00	4778.17
β_{first}	-1.46	0.44	-2.16	-0.76	1.00	4794.43
β_{repeated}	-0.19	0.01	-0.21	-0.17	1.00	2433.58
$\beta_{\text{Age\_missing}}$	0.01	0.99	-1.58	1.60	1.00	6351.09
$\beta_{\text{Age\_75}}$	3.68	0.25	3.28	4.09	1.00	1222.88
$\beta_{\text{Age\_50}}$	1.53	0.13	1.33	1.72	1.00	1110.39
$\beta_{\text{Age\_25}}$	1.25	0.13	1.05	1.45	1.00	1182.94
β_{Iw}	4.94	0.15	-5.18	-4.69	1.01	442.12
β_{Fw}	-4.68	0.15	4.43	4.91	1.01	436.35
σ	1.01	1.01	0.05	2.98	1.00	4236.47
$\rho_{\text{FwM[2,1]}}$	0.93	0.00	0.93	0.93	1.01	599.97
$\sigma_{\text{FwM[1]}}$	1.00	0.00	0.99	1.01	1.00	4083.05
$\sigma_{\text{FwM[2]}}$	1.07	0.01	1.06	1.07	1.00	3887.48

7 Discussion

Our investigation focused on two critical objectives: predicting factors associated with diet completion and evaluating the efficacy of missing data imputations. We approached this in two phases: initially analyzing the entire dataset and subsequently examining each phase individually.

Weight indeed appears to have the most significant impact on completion, which aligns with expectations given the context of a weight loss program. The imputation process plays a crucial role in allowing us to utilize data from both observed and unobserved variables, thereby providing a more comprehensive understanding of completion rates. Specifically, initial weight demonstrates a positive impact on completion, suggesting that individuals with higher initial weights tend to be more motivated to complete the diet program. Conversely, final weight exhibits a negative impact, indicating that individuals with higher final weights are more likely to drop out of the program. This finding underscores the importance of effectively managing weight throughout the program to enhance retention and overall success.

Consistent with findings from literature reviews and descriptive statistics, age emerged as a significant predictor across all models. Older individuals tend to exhibit higher rates of diet

completion compared to their younger counterparts. This trend aligns with the notion that older individuals typically lead more stable lives, both temporally and financially, allowing for greater dedication and investment in such programs. Additionally, older individuals may have heightened concerns regarding weight-related health issues, given their age.

Interestingly, gender did not appear to significantly influence diet completion rates across all models. Despite a higher proportion of women participating, there were no notable differences in completion rates observed in the descriptive statistics, a trend confirmed by our models

Another variable demonstrating significant variation was `brand_adhesion_type`, evident not only across the entire diet program but also within individual phases. This variability underscores the importance of tailoring program offerings to align with the unique characteristics of each phase, optimizing efficiency and effectiveness.

8 Conclusion

The results presented in this study offer valuable insights not only into attrition within this specific weight loss program but also into a less common approach for predicting attrition, utilizing Bayesian imputation models.

We opted for this technique to avoid discarding cases with missing values since, in our context, these missing values contain valuable and pertinent data that can provide insights into program completion. Instead, we delved into the causal mechanisms behind the missing process, seeking to understand the factors leading to these values being missing and subsequently proceeded to impute estimates.

Our analysis extensively explored completion throughout the entire diet program, encompassing numerical, categorical, and correlation measures across various models. However, it's important to note that our examination of individual phases serves as supplementary analysis to complement the main findings. By focusing on each phase separately, we not only exclude patients who did not participate in specific phases but also disregard historical data for patients involved in those phases.

The study revealed that age plays a significant role in program completion, aligning with existing literature on the subject. Our findings indicate that older individuals are more likely to complete the program compared to younger counterparts, a trend observed not only across the entire diet plan but also within each phase. Older participants exhibit higher completion rates across all three phases when contrasted with younger individuals.

In most instances, previous attempts at completing the diet—such as clients repeating the program—appear to have a negative impact on completion rates. This trend persists across Phase 3 and Phase 2. However, Phase 1 presents a different pattern, with repeat participants demonstrating a higher likelihood of completing Phase 1. This observation aligns with expectations, considering Phase 1's status as the initial phase, likely motivating participants more and being the shortest phase.

The primary limitations of this study arise from the considerable computational resources and time required by the models employed. These constraints necessitated the use of random samples rather than analyzing the entire dataset. Additionally, not all potential variables present in the dataset were thoroughly explored, potentially limiting the depth of analysis and overlooking crucial factors influencing completion rates. Furthermore, particularly in the analysis of the entire diet, we encountered a dataset imbalance as we endeavored to replicate the same completion ratio across our stratified samples. This imbalance may have posed challenges for our imputation and prediction processes.

In conclusion, despite the identified limitations, our study marks a significant advancement in understanding attrition within weight loss programs. The model we have developed and applied not only provides valuable insights into completion rates but also lays the groundwork for future research endeavors. While our current study utilized random samples and focused on specific variables, it sets the stage for a larger, more comprehensive investigation. Our approach, incorporating Bayesian imputation models and exploring causal mechanisms behind missing data, offers a robust framework that can be extended to broader contexts with missing censored data, incorporating covariates for enhanced predictive power. Future studies could explore incorporating Bayesian variable selection techniques to further refine the model and uncover additional insights. Overall, our study serves as a solid foundation for future research efforts seeking to delve deeper into the complexities of program completion and attrition prediction in weight loss interventions.

References

Austin, P. C. and Escobar, M. D. (2005). Bayesian modeling of missing data in clinical research. *Computational Statistics Data Analysis*, 49(3):821–836.

Bitner, M. J., Faranda, W. T., Hubbert, A. R., and A., Z. V. (1997). Customer contributions and roles in service delivery. *International Journal of Service Industry Management*, 8(3):193–205.

Buckinx, W. and Van den Poel, D. (2010). Customer base analysis: partial defection of behaviourally loyal clients in a non-contractual fmcg retail setting. *European Journal of Operational Research*, 164(1):252–268.

Bush, R., Vemulakonda, V., Corbett, S., and Chiang, G. (2014). Can we predict a national profile of non-attendance paediatric urology patients: a multi-institutional electronic health record study. *Inform Prim Care*, 21(3):132–138.

Carreras-García, D., Delgado-Gómez, D., Llorente-Fernández, F., and Arribas-Gil, A. (2020). Patient no-show prediction: A systematic literature review. *Entropy*, 22(6).

Dashtban, M. and Li, W. (2022). Predicting non-attendance in hospital outpatient appointments using deep learning approach. *Health Systems*, 11(3):189–210.

Fabricatore, A. N., Wadden, T. A., Moore, R. H., Butryn, M. L., Heymsfield, S. B., and Nguyen, A. M. (2009). Predictors of attrition and weight loss success: Results from a randomized controlled trial. *Behaviour Research and Therapy*, 47(8):685–691.

Fayyaz, H., Phan, T.-L. T., Bunnell, H. T., and Beheshti, R. (2022). Who will leave a pediatric weight management program and when? – a machine learning approach for predicting attrition patterns. *CoRR*, abs/2202.01765.

Felderer, M. and Travassos, G. H. (2020). *Contemporary Empirical Methods in Software Engineering*. Springer Nature Switzerland.

Gelman, Andrew Carlin, J. B. . S. H. S. . D. D. B. . V. A. . R. D. B. (2013). *Bayesian Data Analysis*. Chapman and Hall/CRC, 3 edition.

Hayton, C., Clark, A., Olive, S., Browne, P., Galey, P., Knights, E., Staunton, L., Jones, A., Coombes, E., and Wilson, A. M. (2013). Barriers to pulmonary rehabilitation: Characteristics that predict patient attendance and adherence. *Respiratory Medicin*, 107(3):401–407.

Honas, J. J., Early, J. L., Frederickson, D. D., and O'Brien, M. S. (2003). Predictors of attrition in a large clinic-based weight-loss program. *The official journal of The Obesity Society*, 11(7):809–917.

Kearney, A., Rosala-Hallas, A., Bacon, N., Daykin, A., Shaw, A. R. G., Lane, A. J., Blazeby, J. M., Clarke, M., Williamson, P. R., and Gamble, C. (2018). Reducing attrition within clinical trials: The communication of retention and withdrawal within patient information leaflets. *PLOS ONE*, 13(10):1–16.

Kumamaru, H., Lee, M. P., Choudhry, N. K., Dong, Y.-H., Krumme, A. A., Khan, N., Brill, G., Kohsaka, S., Miyata, H., Schneeweiss, S., and Gagne, J. J. (2018). Using previous medication adherence to predict future adherence. *Journal of Managed Care & Specialty Pharmacy*, 24(11):1146–1155.

McElreath, R. (2023). *Statistical Rethinking - A Bayesian Course with Examples in R and Stan*. CRC Press, 2 edition.

Moroshko, J., Brennan, L., and O'Brien, P. (2011). Predictors of dropout in weight loss interventions:a systematic review of the literature. *Obesity Reviews*, 12(11):912–934.

Robins, J. M., Rotnitzky, A., and Zhao, L. P. (1994). Estimation of regression coefficients when some regressors are not always observed. *Journal of the American Statistical Association*, 89(427):846–866.

Rubin, D. B. (2018). *Flexible Imputation of Missing Data*. Chapman and Hall/CRC, 2 edition.

Tamaddoni Jahromi, A., Stakhovych, S., and Ewing, M. (2014). Managing b2b customer churn, retention and profitability. *Industrial Marketing Management*, 43(7):1258–1268.

Teixeira, P., Going, S., Houtkooper, L., Cussler, E., Metcalfe, L., Blew, R., Sardinha, L., and Lohman, T. (2004). Pretreatment predictors of attrition and successful weight management in women. *International Journal of Obesity*, 28:1124–1133.

van de Schoot, R., Depaoli, S., Gelman, A., King, R., Kramer, B., Märtens, K., Tadesse, M. G., Vannucci, M., Willemsen, J., and Yau, C. (2021). Bayesian statistics and modelling. *Journal Nature Reviews Methods Primers*, 1(3).

Appendix

Figure 13: Trace Plot of the Markov chains for Phase 1 Model

Note: The figure displays Trace plot of the Markov chain from the final model of the whole diet. The gray region is warmup, during which the Markov chain was adapting to improve sampling efficiency. The white region contains the samples used for inference.

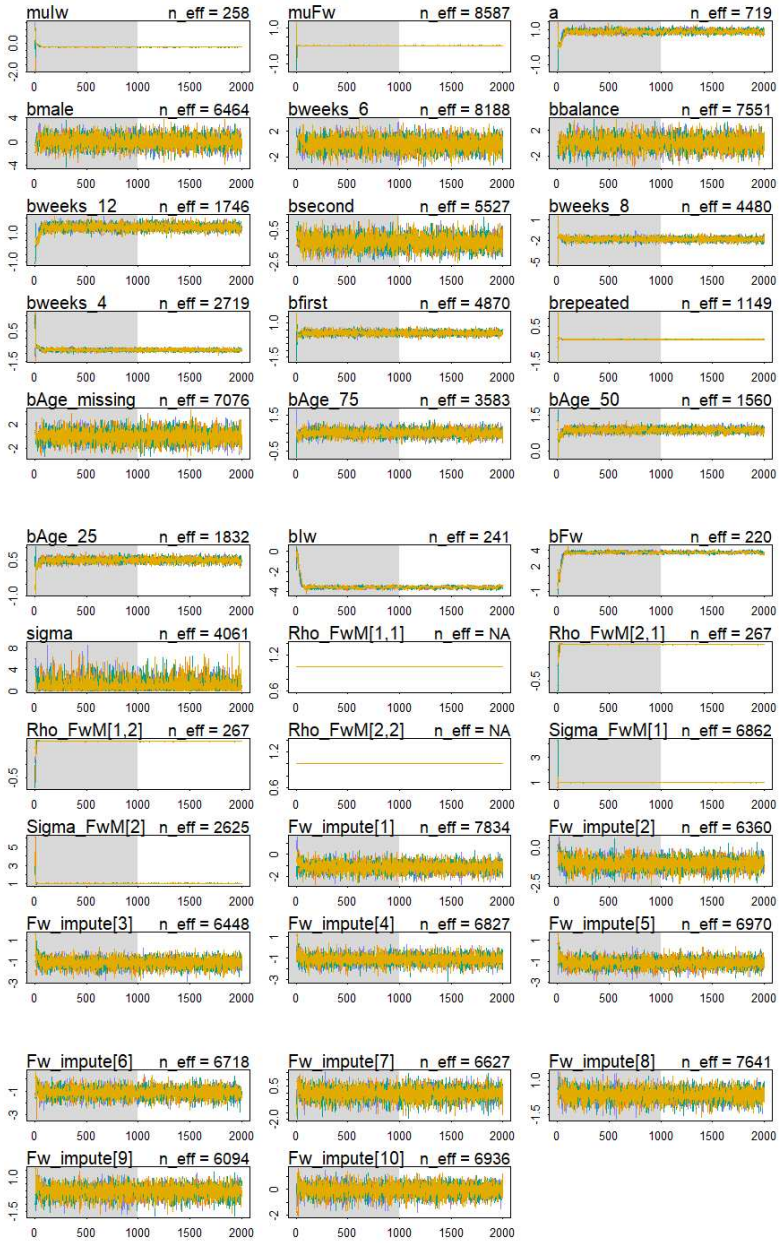


Figure 14: Trace Plot of the Markov chains for Phase 2 Model

Note: The figure displays Trace plot of the Markov chain from the final model of the whole diet. The gray region is warmup, during which the Markov chain was adapting to improve sampling efficiency. The white region contains the samples used for inference.

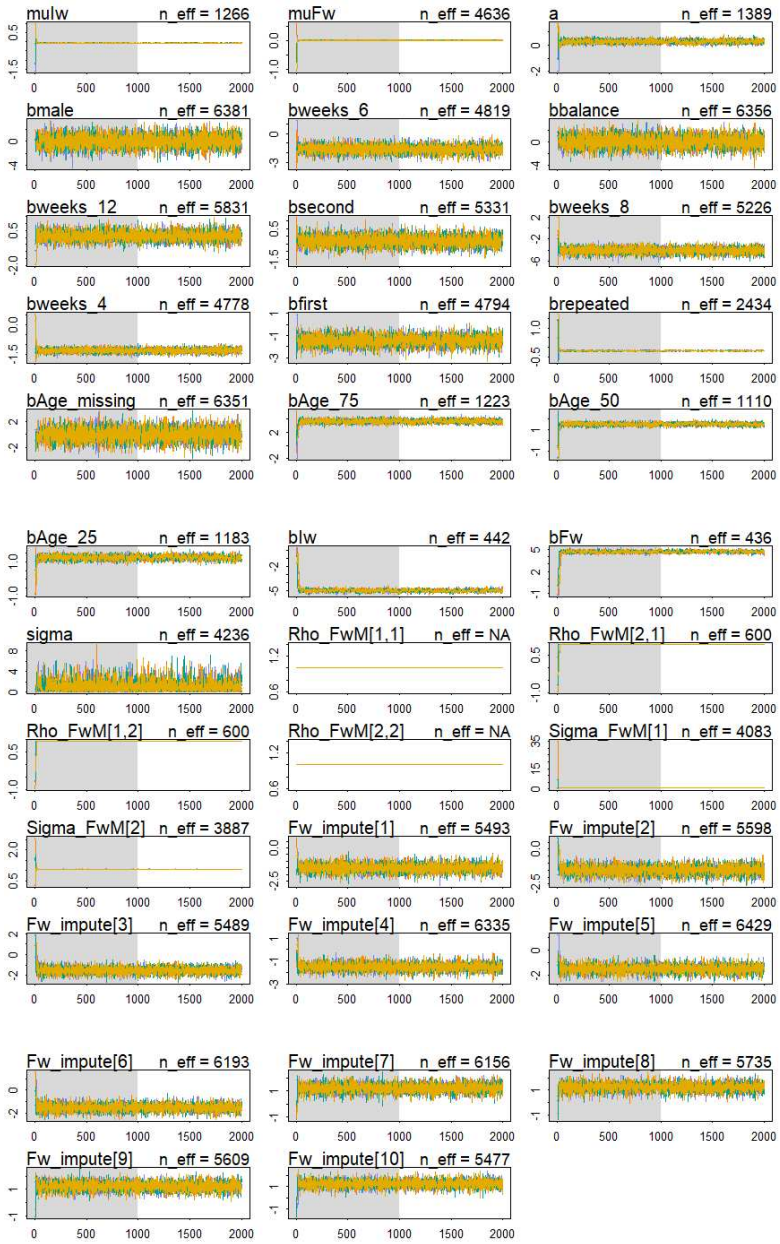


Figure 15: Trace Plot of the Markov chains for Phase 3 Model

Note: The figure displays Trace plot of the Markov chain from the final model of the whole diet. The gray region is warmup, during which the Markov chain was adapting to improve sampling efficiency. The white region contains the samples used for inference.