Journal of Artificial Intelligence and Data Science (JAIDA)
e-ISSN: 2791-8335
https://dergipark.org.tr/pub/jaida

Vol. 4, No. 1, pp. 1-10, 2024
Research Article
Received: 15/05/2024 Accepted: 24/06/2024

Predictive Modeling of Endovenous Laser Ablation Treatment Outcome in Varicose Veins

Steve CHUNG [1,*], Serin ZHANG [2], Sanjay SRIVATSA[3]

Abstract

Varicose veins afflict a significant portion of adults, with approximately 30% experiencing this condition, which often necessitates medical treatment like endovenous laser ablation (EVLA). EVLA has emerged as a highly effective and minimally invasive treatment. However, despite its efficacy, there is a lack of literature on predictive modeling of EVLA treatment outcomes considering both surgical settings and patient characteristics. In this study, we present a comprehensive analysis employing logistic regression under both maximum likelihood (ML) and Bayesian frameworks, as well as support vector machine (SVM) regression. Our results indicate that Bayesian logistic regression with uniform prior demonstrates superior performance. Furthermore, through repeated random sub-sampling validation, we confirm the robustness of our models in predicting successful EVLA outcomes. These findings provide the potential of machine learning techniques in augmenting predictive capabilities in medical decision-making. Our study contributes to the burgeoning literature on predictive modeling in medical contexts, offering insights into the optimization of EVLA treatment outcomes.

Keywords: Bayesian logistic regression; Endovenous laser ablation; Logistic regression; Support vector machine; Varicose veins.

1. Introduction

Varicose veins, although often considered a cosmetic concern, can lead to serious complications such as blood clots and leg ulcers if left untreated, emphasizing the importance of research and intervention. Varicose veins occur with the breakdown of the veins one-way valves causing the vein to dilate. The most common treatment for varicose veins is endovenous laser ablation (EVLA). EVLA has been proven to be the most effective and least invasive treatment for varicose veins. This procedure is done by inserting a laser fiber into the vein, typically near the knee or ankle. The laser is then retracted through vein, causing the vein to occlude. According to Mundy [6], EVLA treatment success rates have been over 89%. EVLA was approved by the National Institute for Health and Clinical Excellence (NICE) in March 2004. This procedure has become very common since it avoids the complications of open surgeries while still having spectacular results which is one factor that is often considered when judging the overall excellence of the surgery. Figure 1 provides an illustration of EVLA treatment.

Although the success rate of EVLA treatment is fairly high, there hasn't been any literature on modeling the EVLA based on both the settings of the surgery and patients' conditions. Some have examined the surgery settings has predictors of the success rate of EVLA treatment. Mordon [7] considered a mathematical model that was produced to determine significant variables in EVLA treatment and optimal levels of endovenous laser Treatment variables that result in minimal vein damage and side effects. They considered both the pulsed and continuous mode of the laser treatment. For each mode, the model determined the optimal linear endovenous energy density

*Corresponding author

Steve CHUNG*; California State University Fresno, Mathematics Department, USA; e-mail: schung@csufresno.edu; 0000-0001-7255-7244

Serin ZHANG; California State University Fresno, Information Systems and Decision Sciences Department, USA; e-mail: serin20@gmail.com; 0009-0003-8745-9534

Sanjay SRIVATSA; Heart Artery Vein Center of Fresno, USA; e-mail: drsanjay@gmail.com; 0000-0002-8661-3187

(LEED) (J/cm) for both 3mm and 5mm vein diameter. They also concluded that pullback distance and laser wavelength do not significantly affect treatment outcome.

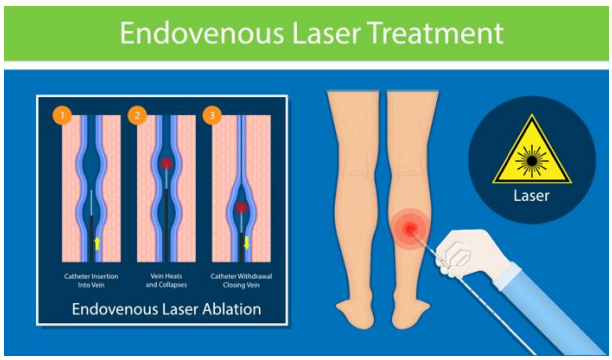

Figure 1. *An illustration of EVLA.*

While the success rate of EVLA treatment is notably high, there remains a gap in the literature regarding the modeling of EVLA outcomes based on both surgical settings and patients' conditions. Some researchers have delved into the predictive value of surgical settings as indicators of EVLA treatment success. For instance, in [7], a mathematical model was developed to identify significant variables in EVLA treatment and determine optimal levels of endovenous laser treatment variables to minimize vein damage and associated side effects. The study explored both pulsed and continuous modes of laser treatment, aiming to establish the optimal linear endovenous energy density (LEED) (J/cm) for vein diameters of 3mm and 5mm. Notably, the researchers concluded that variables such as pullback distance and laser wavelength did not exert a significant impact on treatment outcomes. Cowpland [3] have examined the factors affecting optimal linear endovenous energy density for endovenous laser ablation in incompetent lower limb veins. The findings indicate that the ideal LEED for endovenous laser ablation of the great saphenous vein lies between 80 J/cm and 100 J/cm to achieve optimal closure rates while minimizing side effects and complications. Longer wavelengths, which target water, may have a lower optimal LEED. Conversely, a LEED below 60 J/cm shows reduced efficacy regardless of the wavelength used.

In this study, our initial focus is on modeling the relationship between EVLA treatment outcomes and both surgical settings and patient characteristics using logistic regression under the frequentist approach. We aim to elucidate the factors influencing the success or failure of EVLA procedures. Following this and more importantly, we compare the predictive capabilities of this logistic regression model with Bayesian logistic regression and support vector machine (SVM) regression. These modeling techniques have been extensively utilized in the medical field to predict the success rates of various treatments. For instance, Yussuff [11] employed logistic regression to predict breast cancer based on mammogram results, identifying mass, architectural distortion, skin thickening, and calcification detection as significant predictors. Similarly, Chadwick [1] utilized univariate and multivariate logistic regression analyses to differentiate between dengue fever and other febrile illnesses, achieving a sensitivity of 74% and a specificity of 79%. More recently, Srivatsa [9] investigated the relative contributions of power output, linear endovenous energy density, and pullback rate using logistic regression, highlighting the significance of power output and LEED.

Moreover, researchers have explored Bayesian and SVM approaches in medical prediction tasks. Zhou [12] introduced a Bayesian approach to identifying important genes in cancer classification, while Riaz [8] developed an adaptive SVM regression model to predict the motion of lung tumors, demonstrating its superiority in accuracy compared to traditional methods. Verplankcke [10] investigated the use of SVM models in predicting mortality of critically ill patients with hematological malignancies, finding comparable predictive power to multiple linear regression. Furthermore, Cheng [2] utilized SVM incorporating protein structure and sequence information to predict changes in protein stability following single amino acid mutations. Additionally, Gelman [4] compared four regression methods, including SVM, in tracking lung tumors, with the artificial neural network regression performing slightly better in terms of mean tracking error.

These studies collectively demonstrate the utility of logistic regression, Bayesian logistic regression, and SVM regression in medical prediction tasks, providing valuable insights into treatment outcomes and disease prognosis. By leveraging these diverse modeling techniques, our study aims to enhance the predictive accuracy of EVLA treatment outcomes, contributing to improved patient care and clinical decision-making.

The paper is organized as follows. Section 2 provides a discussion on the data and methodology. It gives the background of the dataset and it provides methodology describing logistic regression, Bayesian logistic

regression, and support vector machine. In Section 3, we provide the results from the models. Section 4 concludes.

2. Data and Methodology

2.1 Dataset

The dataset was obtained from the Heart, Artery, and Vein Center of Fresno, a local clinic in Fresno, CA. It contains information on 359 veins treated using EVLA. Institutional review board (IRB) approval was not needed since the data had already been collected prior to the study. The clinic obtained all patients' consents, and all patient identifications were masked and not revealed to the researchers. Due to missing observations, the complete case data consists of 272 observations from 2015 to 2017. Table 1 presents the descriptions of the variables and Table 2 provides the summary statistics for all the variables.

Table 1. *Variable definitions and summary statistics.*

Variables	**Definitions**	Min	Q1	Median	Mean	Q1	Max	Std. Dev
Age	age in years	26	56.75	65	63.51	72	91	12.41
Height	height in inches	53	63	65.5	65.58	68	74	3.94
Weight	weight in lbs	100	163.8	186	201.9	234	460	60.25
BMI	body mass index	16.8	27.4	30.85	32.75	36.7	62.4	8.33
Power	power setting in watts	6	9	10	9.871	12	12	1.76
Time	time of ablation in seconds	9	58	94	97.95	130	258	49.6
Length	length of treated vein in cm	2.5	28	45	42.68	56	92	17.8
Energy	energy used in joules	11	527	839	968.1	1318	2581	588.4
LEED	linear endovenous energy density in J/cm	1	17.75	21.5	21.91	27	41	7.15
Pullback	pullback rate in mm/s	2.756	3.781	4.485	4.695	5.252	14.655	1.38

Table 2. *Variable definitions and frequencies.*

Variables	Definitions	Yes	No	Total
DM	dementia	107	165	272
HTN	hypertension	195	77	272
Hyperlip	hyperlipidemia	184	88	272
Renal	renal disease	20	252	272
CHD	congenital heart defects	33	239	272

2.2 Methodology

In this section, we present the methods and models that we have used in our study. In the frequentist approach, we assume a specific probability distribution such as a normal distribution and then, estimate the parameters in the model. Maximum likelihood estimation has been one of the most popular estimation methods in a frequentist setting due to its efficiency, consistency, and asymptotic normality. In the Bayesian approach, estimation of parameters involves treating them as random variables rather than fixed quantities, allowing for the incorporation of prior knowledge and uncertainty into the modeling process. Unlike the frequentist approach, which relies solely on observed data to estimate parameters, Bayesian inference combines prior beliefs about the parameters with likelihood functions derived from the data to obtain posterior distributions. These posterior distributions represent updated beliefs about the parameters after observing the data, reflecting both the information contained in the data and the prior knowledge. Bayesian data analysis has become a well-established component of modern applied statistics and machine learning terminology. However, there is no universal consensus on which approach provides better results. This study aims to compare these approaches, along with a machine learning technique, support vector machine, to find an optimal predictive model for EVLA outcome.

2.2.1 Logistic Regression

Logistic regression is a most commonly used model for predicting a binary response variable. Let Y_i be the binary response variable with the conditional probability $p_i = P(Y_i = 1|X = x)$ where the event $\{Y_i = 1\}$ denotes the success of an outcome for the i-th observation. The logistic regression model has the form

$$\log\left(\frac{p_i}{1-p_i}\right) = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_1 x_{ik}. \quad (1)$$

Solving for p_i from above gives

$$p_i = \frac{1}{1+\exp(\beta_0 + \beta_1 x_{i1} + \cdots + \beta_1 x_{ik})}, \quad (2)$$

which is the probability of the success after observing $x_1, x_2, \ldots, x_k$. The parameters $\beta_0, \beta_1, \ldots, \beta_k$ can be estimated from the maximum likelihood estimation. That is, since *Y* is an independent binary random variable the likelihood function is defined as

$$L(\beta|y) = \prod_i^n \frac{\exp(y_i(\beta_0+\beta_1 x_{i1}+\cdots+\beta_k x_{ik}))}{1+\exp(\beta_0+\beta_1 x_{i1}+\cdots+\beta_k x_{ik})^{n_i}}, \quad (3)$$

where n_i is the total number of *i*-th trial. Differentiating above equation with respect to β gives k+1 equations and solving for β gives the estimated parameters. However, solving this system of nonlinear equations is not easy since the solution cannot be derived algebraically. A numerical method such as Newton's method is often used to obtain the solution.

2.2.2 Bayesian Logistic Regression

In Bayesian analysis, the posterior distribution $p(\beta|y)$ is obtained from the likelihood function $L(\beta|y)$ and a prior distribution $p(\beta)$. That is,

$$p(\beta|y) = \frac{L(\beta|y)p(\beta)}{\int L(\beta|y)p(\beta)d\beta}. \quad (4)$$

Choosing an appropriate prior distribution is critical in Bayesian setting because the posterior heavily depends on it. For instance, if the prior distribution is chosen to be a beta distribution, then it can be easily be shown that the posterior distribution belongs to a class of beta distributions. Of course, a beta prior distribution may be subjective since the parameter space for β lie in the whole real, whereas a beta distribution only takes the values in (0,1). On the other hand, if the prior distribution is chosen as a normal then there is no closed form for the posterior distribution (unless the likelihood function is normal) and hence, sampling from this posterior is not easy.

For simplicity, we assume that the parameters $\beta_0, \beta_1, \ldots, \beta_k$ are independent. We used two prior distributions in our work. They are

$$\beta_j \sim N(\mu, \sigma^2) : p(\beta_j) = \frac{1}{\sigma\sqrt{2\pi}} exp\left(\frac{-(\beta_j-\mu)\hat{}2}{2\sigma^2}\right) \text{ and } \beta_j \sim Unif(a,b): p(\beta_j) = 1. \quad (5)$$

A normal prior puts a heavy weight near μ while the uniform prior assigns equal weight and hence, it is called a non-informative prior. As mentioned previously, sampling from a posterior is difficult in many cases. For instance, under a normal prior distribution, the posterior is

$$p(\beta|y) \propto \prod_i^n \frac{\exp(y_i(\beta_0+\beta_1 x_{i1}+\cdots+\beta_k x_{ik}))}{1+\exp(\beta_0+\beta_1 x_{i1}+\cdots+\beta_k x_{ik})^{n_i}} \; exp\left(\frac{-\sum(\beta_j-\mu)^2}{2\sigma^2}\right). \quad (6)$$

However, there is no closed form for this posterior distribution and sampling and making inference from this posterior is not straightforward. Therefore, we resort to the random walk metropolis algorithm which is ubiquitous tool for producing dependent simulations from an arbitrary distribution. The reader is referred to [4] for details on this algorithm. In the Appendix, trace plots and density plots are shown to verify the convergence of the algorithm for each parameter β_j.

2.2.3 Support Vector Machine Regression

Support Vector Machine (SVM) regression is a supervised learning algorithm used for regression tasks, where the goal is to predict outcomes. Unlike traditional regression methods that minimize error directly, SVM regression aims to fit a "tube" around the data points, with the goal of including as many points as possible within the tube while minimizing the margin violations (points outside the tube). SVM regression aims to find a hyperplane that best fits the data points while maximizing the margin, subject to a tolerance ϵ. This hyperplane is used to predict the target values for new data points. The objective of SVM regression is to minimize the following function:

$$\frac{1}{2}\left|\left|w\right|\right|^2 + C\sum_{i=1}^{n}(|y_i - w\,\phi(x_i) - b| - \epsilon)_+, \tag{7}$$

where *w* is the weight, *b* is the bias term, ϵ is the tube radius (tolerance), *C* is the regularization parameter (trade-off between maximizing the margin and minimizing the errors. The optimization problem involves finding the optimal values for *w* and *b* that minimize the objective function while satisfying the margin constraints. This is be formulated as a quadratic programming problem and solved using optimization technique.

3. Results

The primary objective of this paper is to identify the optimal model for predicting successful outcomes of EVLA treatment. To achieve this goal, we explore frequentist, Bayesian, and machine learning approaches. Our aim is to determine the most effective approach, laying the groundwork for the development of superior models in the future. Model assessment relies on repeated random sub-sampling validation, commonly referred to as Monte Carlo cross-validation. We randomly partition the dataset into four subsamples, with three utilized as training data and the remaining subsample serving as validation data for testing the model. This process is repeated 50 times to ensure robust evaluation.

Given that this is a binary classification problem, each case in the validation set is classified as either a correct or incorrect prediction. We define success as $Y_i = 1$ if the estimated probability is greater than or equal to 0.90. This criterion is referred to as accuracy (ACC) in our results. Sensitivity (TPR), specificity (TNR), and precision were calculated based on the following formulas.

$$Accuracy = ACC = \frac{TP+TN}{TP+TN+FP+FN} \tag{8}$$

$$Sensitivity = TPR = \frac{TP}{TP+FN} \tag{9}$$

$$Specificity = TNR = \frac{TN}{TN+FP} \tag{10}$$

$$Precision = \frac{TP}{TP+FP} \tag{11}$$

where TP, FP, TN, and FN represent the number of true positives, false positives, true negatives, and false negatives.

First, we fitted a logistic regression to the whole dataset using the maximum likelihood estimation method as presented earlier. This approach provides parameter estimates that can be easily interpreted, along with the significance of each input variable in our model. Table 3 provides the results of this frequentist approach. Power emerges as the most significant variable, along with Energy and LEED. This result is consistent with the findings of [9].

Table 3. *Logistic regression analysis output.*

Variable	Estimate	Std. Error	Z value	P-value	Significance
Intercept	24.309	14.407	1.687	0.092	.
Age	-0.001	0.020	-0.049	0.961	
Height	-0.305	0.217	-1.407	0.159	
Weight	0.056	0.034	1.639	0.101	
BMI	-0.385	0.209	-1.836	0.066	.
Power	0.576	0.158	3.641	0.000	***
Time	-0.019	0.017	-1.083	0.279	
Length	-0.083	0.043	-1.943	0.052	.
Energy	0.006	0.002	2.558	0.011	*
LEED	-0.222	0.107	-2.081	0.037	*
Pullback	-0.385	0.244	-1.576	0.115	
DM	0.132	0.478	0.275	0.783	
HTN	-0.322	0.550	-0.584	0.559	
Hyperlip	0.729	0.556	1.311	0.190	
Renal Disease	-0.685	0.753	-0.910	0.363	
CHD	1.041	0.726	1.435	0.151	

Table 4. *Logistic regression analysis from the backward selection process.*

Variable	Estimate	Std. Error	Z value	P-value	Significance
Intercept	21.434	13.240	1.619	0.105	
Height	-0.311	0.200	-1.551	0.121	
Weight	0.055	0.031	1.801	0.072	.
BMI	-0.379	0.190	-1.997	0.046	*
Power	0.497	0.147	3.383	0.001	***
Length	-0.080	0.037	-2.192	0.028	*
Energy	0.004	0.002	2.088	0.037	*
LEED	-0.120	0.083	-1.45	0.147	
CHD	1.045	0.702	1.489	0.136	

Subsequently, we applied backward stepwise logistic regression to fit a subset model, selecting the variables Height, Weight, BMI, Power, Length, Energy, LEED, and CHD for classification purposes. The results, displayed in Table 4, indicate that Power is the most significant variable, alongside BMI, Length, and Energy. In contrast, LEED is non-significant in this model, possibly due to its correlation with other variables such as Power and Length.

In Figure 2, a ROC curve displays the Area Under the Curve (AUC) derived from one of the 50 Monte Carlo cross-validations conducted. Table 5 presents the average accuracy, sensitivity, specificity, precision, and AUC across all 50 cross-validations, with the values in parentheses indicating the corresponding standard deviations. Concerning accuracy, sensitivity, specificity, and precision, the Bayesian model with a uniform prior exhibits slightly superior performance compared to (frequentist) logistic regression. Moreover, the Bayesian model with a normal prior outperforms both logistic regression and the Bayesian model with a uniform prior, albeit marginally.

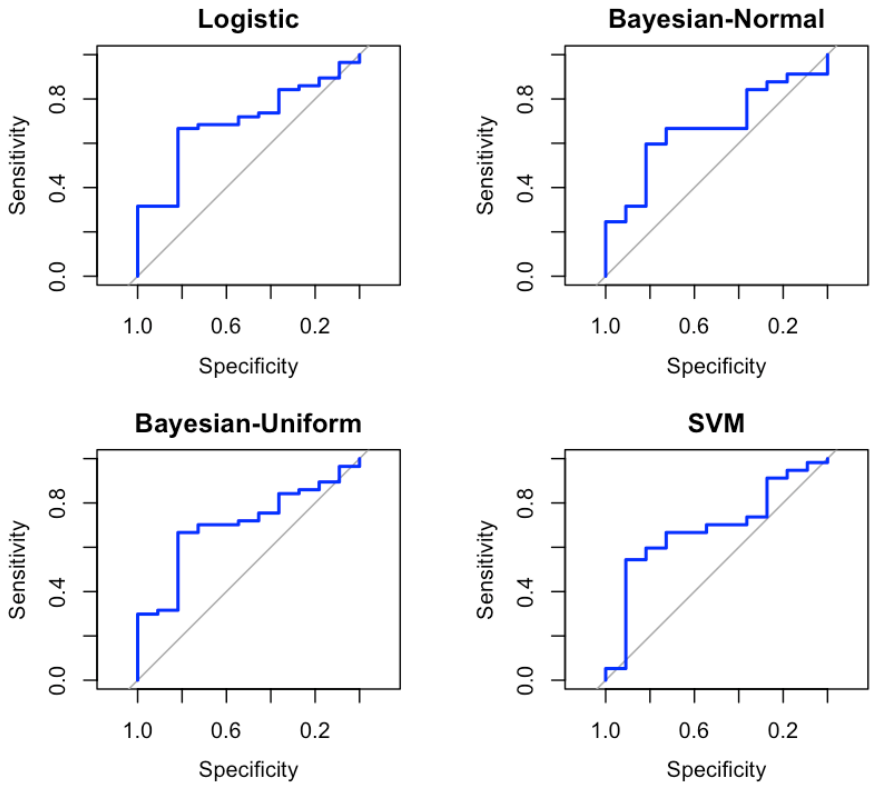

Figure 2. *A sample ROC curve showing the area under the curve (AUC).*

Table 5. *A comparison of different models using accuracy, sensitivity, specificity, precision, and AUC.*

	Logistic	Bayesian (Normal)	Bayesian (Uniform)	SVM
Accuracy	0.861	0.856	0.862	0.816
	(0.040)	(0.038)	(0.039)	(0.036)
Sensitivity	0.102	0.079	0.113	0.074
	(0.092)	(0.100)	(0.102)	(0.088)
Specificity	0.984	0.983	0.985	0.937
	(0.018)	(0.019)	(0.019)	(0.028)
Precision	0.579	0.446	0.585	0.150
	(0.401)	(0.424)	(0.396)	(0.159)
AUC	0.687	0.688	0.687	0.583
	(0.066)	(0.070)	(0.068)	(0.070)

4. Conclusions

Based on the analysis conducted, several key findings emerge regarding the predictive modeling of endovenous laser ablation (EVLA) treatment outcomes:

(a) Logistic Regression Analysis: Initial logistic regression analysis identified significant predictors for EVLA outcomes, notably power, energy, and linear endovenous energy density (LEED). This finding aligns with previous research that highlights the importance of these variables in treatment success.

(b) Comparison of Modeling Approaches: The study compared three modeling approaches: logistic regression, Bayesian logistic regression, and support vector machine (SVM) regression. Across these methods, Bayesian logistic regression with a uniform prior demonstrated slightly superior performance in terms of accuracy, sensitivity, specificity, precision, and AUC compared to logistic regression and Bayesian with a normal prior. SVM regression, while providing acceptable accuracy, showed comparatively lower performance in terms of sensitivity and precision.

(c) Clinical Implications: The findings suggest that Bayesian logistic regression, particularly with a uniform prior, may offer improved predictive capabilities for EVLA treatment outcomes compared to traditional logistic regression. This insight can inform clinical decision-making by providing clinicians with a more accurate assessment of the likelihood of treatment success.
(d) Limitations and Future Directions: While Bayesian logistic regression shows promise, further research is warranted to validate and refine the model. Additionally, exploring additional variables or incorporating advanced machine learning techniques may enhance predictive accuracy further. Furthermore, external validation using data from diverse clinical settings would strengthen the generalizability of the findings.

In conclusion, leveraging Bayesian logistic regression models, particularly with a uniform prior, holds potential for enhancing the prediction of EVLA treatment outcomes. By refining predictive models, clinicians can better tailor treatment strategies, ultimately improving patient care and outcomes in the management of varicose veins.

Declaration of Interest

The authors declare that there is no conflict of interest.

Author Contributions

Steve Chung: the corresponding author, write-up, model fitting, programming. Serin Zhang: write-up, modeling fitting, programming. Sanjay Srivatsa: providing the data, review and editing.

References

[1] D. Chadwick, B. Arch, A. Wilder-Smith, and N. Paton, "Distinguishing dengue fever from other infections on the basis of simple clinical and laboratory features: Application of logistic regression analysis," Journal of Clinical Virology, vol. 35, no. 2, pp. 147–153, 2006.

[2] J. Cheng, A. Randall, and P. Baldi, "Prediction of protein stability changes for single-site mutations using support vector machines," Proteins: Structure, Function, and Bioinformatics, vol. 62, no. 4, pp. 1125–1132, 2006.

[3] C. A. Cowpland, A. L. Cleese, and M. S. Whiteley, "Factors affecting optimal linear endovenous energy density for endovenous laser ablation in incompetent lower limb truncal veins–A review of the clinical evidence," Phlebology, vol. 32, no. 5, pp. 299-306, 2017.

[4] A. Gelman and D. B. Rubin, "Avoiding model selection in Bayesian social research," Sociological Methodology, vol. 25, pp. 165-173, 1995.

[5] T. Lin, L. I. Cervino, X. Tang, N. Vasconcelos, and S. B. Jiang, "Fluoroscopic tumor tracking for image-guided lung cancer radiotherapy," Physics in Medicine and Biology, vol. 54, no. 4, pp. 981–992, 2009.

[6] L. Mundy, T. L. Merlin, R. A. Fitridge, and J. E. Hiller, "Systematic review of endovenous laser treatment for varicose veins," British Journal of Surgery, vol. 92, no. 10, pp. 1189–1194, 2005.

[7] S. Mordon, B. Wassmer, and J. Zemmouri, "Mathematical modeling of 980-nm and 1320-nm endovenous laser treatment," Lasers in Surgery and Medicine, vol. 39, no. 3, pp. 256–265, 2007.

[8] N. Riaz, P. Shanker, R. Wiersma, O. Gudmudsson, W. Mao, B. Widrow, and L. Xing, "Predicting respiratory tumor motion with multi-dimensional adaptive filters and support vector regression," Physics in Medicine and Biology, vol. 54, no. 19, pp. 5735–5748, 2009.

[9] S. S. Srivatsa, S. Chung, and V. Sidhu, "The relative roles of power, linear endovenous energy density, and pullback velocity in determining short-term success after endovenous laser ablation of the truncal saphenous veins," Journal of Vascular Surgery: Venous and Lymphatic Disorders, vol. 7, no. 1, pp. 90-97, 2019.

[10] T. Verplancke, S. Van Looy, D. Benoit, S. Vansteelandt, P. Depuydt, F. De Turck, and J. Decruyenaere, "Support vector machine versus logistic regression modeling for prediction of hospital mortality in critically ill patients with haematological malignancies," BMC Medical Informatics and Decision Making, vol. 8, no. 56, pp. 1–8, 2008.

[11] H. Yussuff, N. Mohamad, U. K. Ngah, and A. S. Yahaya, "Breast cancer analysis using logistic regression," IJRRAS, vol. 10, January 2012.

[12] X. Zhou, K. Y. Liu, and S. T. Wong, "Cancer classification and prediction using logistic regression with Bayesian gene selection," Journal of Biomedical Informatics, vol. 37, no. 4, pp. 249–259, 2004.

Appendix

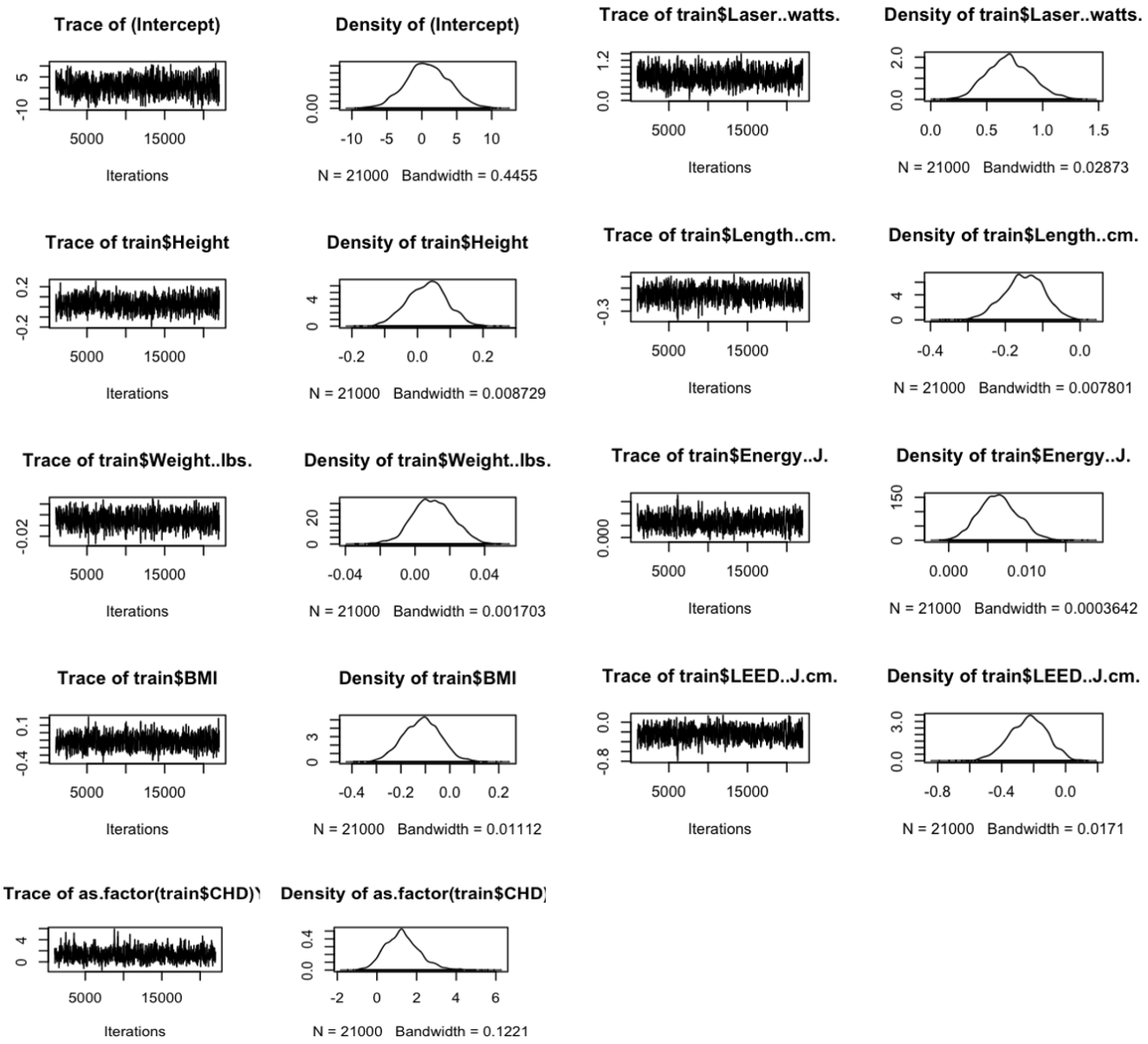

Figure 3. *Trace and density plots of the chains under Bayesian with normal prior.*

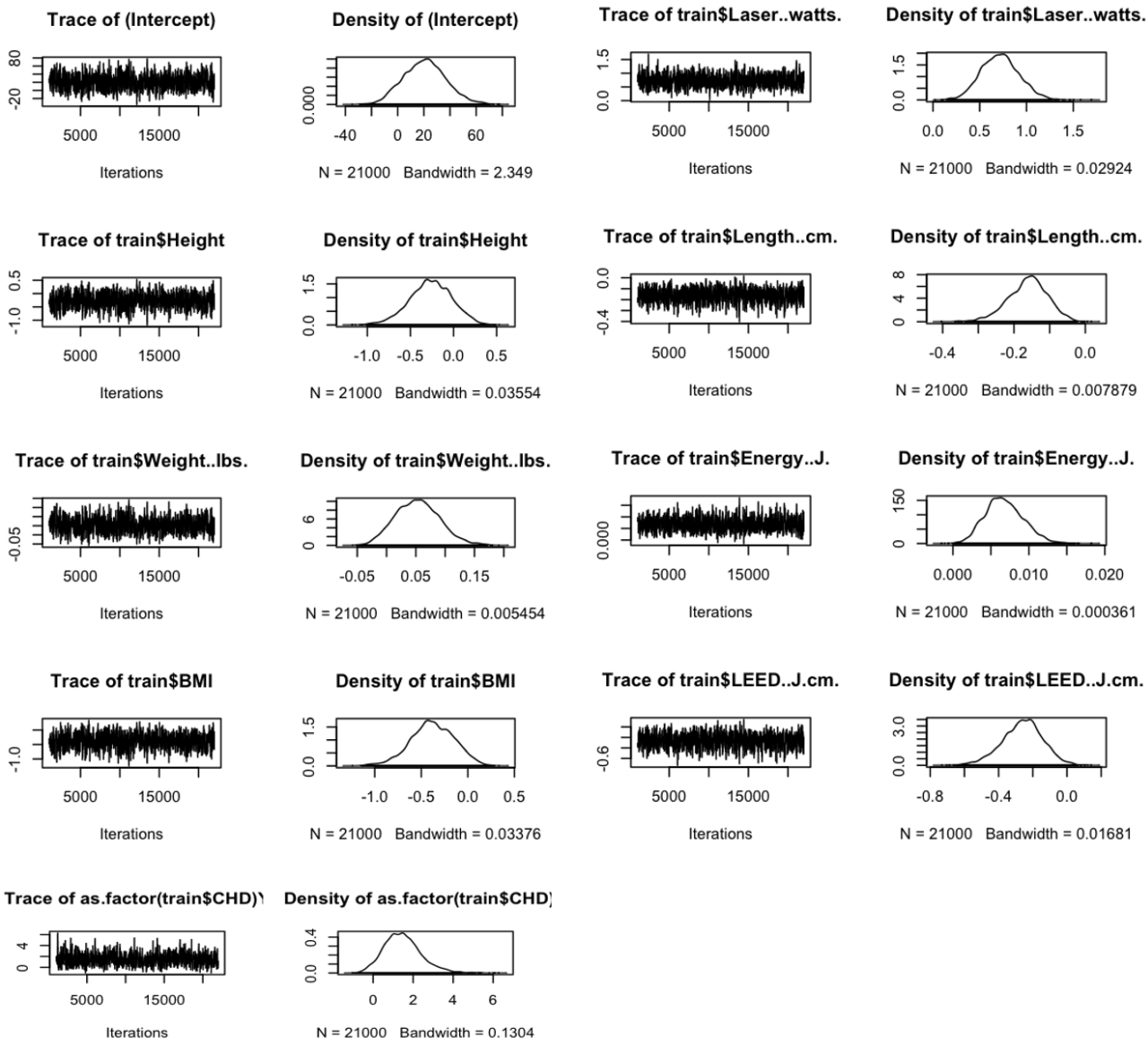

Figure 4. *Trace and density plots of the chains under Bayesian with uniform prior.*