Statistical Learning for Best Practices in Tattoo Removal

Richard P. Yim*†

Project advisors: Jamie Haddock and Deanna Needell**

Abstract. The causes behind complications in laser-assisted tattoo removal are currently not well understood, and in the literature relating to tattoo removal the emphasis on removal treatment is on removal technologies and tools, not best parameters involved in the treatment process. Additionally, the very challenge of determining best practices is difficult given the complexity of interactions between factors that may correlate to these complications. In this paper we apply a battery of classical statistical methods and techniques to identify features that may be closely correlated to causes of complication during the tattoo removal process, and report quantitative evidence for potential best practices. We develop elementary statistical descriptions of tattoo data collected by the largest gang rehabilitation and reentry organization in the world, Homeboy Industries; perform parametric and nonparametric tests of significance; and finally, produce a statistical model explaining treatment parameter interactions, as well as develop a ranking system for treatment parameters utilizing bootstrapping and gradient boosting.

Key words. Laser Assisted Tattoo Removal, Parametric Tests, Non-Parametric Tests, Binary Classification, Logistic Regression, Gradient Boosting, Bootstrapping

1. Introduction.

Current best treatment practices for laser-assisted tattoo removal are ad hoc at best, and the interactions between parameters involved in the tattoo removal treatment process are not well understood. For instance, it is believed that patient demographic factors may potentially affect treatment outcomes (e.g., factors such as patient age and gender), and that these factors are a source of variation in physiological processes between individuals (i.e., skin reactivity and healing time vary demographically). Additionally, much of the research surrounding tattoo removal is focused on the clinical trial setting with an emphasis on discovering new innovations for laser-assisted tattoo removal [1]. While there is considerable debate on laser pulse duration standards, it still remains that there is no precise and standard system for best practices that has been shown to be both optimally effective and safe [15]. A reason for the lack of understanding of safe treatment practices is a general lack of records of treament parameters causing complications [11].

For our study we explore data recorded by Homeboy Industries, the largest gang rehabilitation program in the world, which offers free tattoo removal treatment services for former gang members seeking to remove gang-related tattoos. We utilize various statistical methods and demonstrate some results of inference relating to laser-assisted tattoo removal treatment procedures performed in a practical clinical setting at Homeboy Industries in Los Angeles.

In our data there are four types of complications that are associated with tattoo removals: hyperpigmentation and hypopigmentation, or increased and decreased skin pigmentation, respectively; scarring, visible tissue regrowth; and keloids, scarring with excessive skin overgrowth in the treated area. We emphasize that while the complication rate among the studied

†*Department of Mathematics, University of California, Los Angeles (UCLA), Los Angeles, CA (richyim555@g.ucla.edu).

Copyright © SIAM
Unauthorized reproduction of this article is prohibited

sample is low, the main concern for all tattoo-removal practitioners is a commitment to doing as little harm as possible. For this reason, our primary concern with the available data is to simply understand complication occurrence and nonoccurence in general, grouping all instances of complications together without distinction. The main significance of this design is to understand the tattoo removal process as one that may potentially discourage patients seeking removal from continuing the treatment process, and cause unnecessary grief in addition to any preexisting tattoo regret, as is commonly surveyed amongst individuals with tattoos [21].

With these concerns in mind we broadly study the following research questions:

1. What demographic factors, if any, place a patient at an increased likelihood of experiencing complications?
2. How are clinicians currently approaching the laser tattoo removal process–are there any significant sources of variation with respect to tattoo demographic factors and laser settings (e.g., face tattoos, tattoo age, etc...)?
3. What laser removal treatment parameters seem to be statistically significant with complication occurrence, and at what particular settings?
4. If healthcare practitioners were to be selectively wary of certain treatment parameters in the laser tattoo removal process is there a potential rank of caution in selecting certain treatment parameters?

Generally, in this paper we apply a full range of inference techniques and machine learning methods to develop an understanding of best tattoo removal treatment parameters. The data measures treatment settings used on a tattoo over multiple treatment appointments. Data is provided at both the patient level and the treatment level for a given tattoo. With regards to methods used, we begin with analyzing the significance of patient demographic factors that are understood to be correlated with complication occurrence by utilizing simple statistical tests, addressing the first research question. We then perform parametric and nonparametric tests of significance for analysis of variance (e.g., Kruskal-Wallis, Wilcoxon rank sum and randomization) and use Tukey's honestly significant difference (HSD) test to identify important sources of variation for current clinician approaches to tattoo removal. Finally, we fit a logistic regression model to identify statistically significant treatment parameters, and perform multiple gradient boosting using decision trees to form a ranking of treatment parameters that a health care practitioner should be most aware of during laser-assisted tattoo removal procedures—the aforementioned "rank of caution."

This paper is organized as follows: Section 2 details the dataset and provides descriptions of the methods utilized; Section 3 presents results of the methods employed; Section 4 discusses the results; we conclude the paper in Section 5 with last remarks.

2. Data and Methods.

The individuals that come through Homeboy industries are often former gang members, and many of them have tattoos that may be offensive or attract negative attention. In this section, we discuss the various datasets recorded by Homeboy Industries. The final cleaned datasets that were used in our analysis detailed tattoo level data and patient demographic data.

We also utilized many different statistical tests and methods, both parametric and nonparametric. We detail the properties of these tests and their requirements for legitimate

application to the data. All computations were done with R-4.0.3 [20]. We used the `caret` library as the primary driver for applying the machine learning algorithms in our study. All computing was done on a Linux machine running Ubuntu 20.04 with an Intel 10700K processor.

2.1. Patient Demographic Data. Including observations with missing values, there are a total of 2,118 tattoo observations among 502 patients recorded in the data. We briefly detail the various factors that were recorded at the patient demographic level in Table 1. We note the importance of the Fitzpatrick score—a measure developed in 1975 to classify skin tones [8]—as it is currently believed in the literature that the efficacy of selected treatment parameters for laser-assisted tattoo removal is dependent on the Fitzpatrick score.

Table 1

Table detailing patient demographic and patient-level factors of interest.

Variable	Description
Patient Age	Age of the patient as of 10 June 2020 (integer-valued)
Sex	Male or female (one-hot encoded as male/not-male, binary)
Ethnicity	Hispanic/Latino, or not (one-hot encoded, binary)
Race	Pacific Islander, American/Alaskan Indian, Black, Asian, Latino/Hispanic, white, multiracial, other (nominal)
Treatment Total	Total number of treatment visits at Homeboy for tattoo removal (integer-valued)
Total Tattoos	Total number of tattoos listed by patient (integer-valued)
Fitzpatrick Score	Fairest tone (I) to deeply pigmented (VI) (ordinal)
Complications	Indication of whether a patient ever experienced any complication (one-hot encoded, binary)

2.2. Tattoo Level Data. At the tattoo level, the data provided by Homeboy Industries includes characteristics of tattoos that went through laser assisted tattoo removal. At the tattoo level we perform statistical analysis on various factors grouped by whether the observed tattoo ever experienced a complication in the treatment sequence. We briefly detail the tattoo level features in Table 2. The data at the tattoo level are important because it provides insight into the significance of tattoo composition in whether a complication is likely to develop as a result of undergoing laser treatment as recorded, as well as challenge some assumptions of existing best practices.

2.3. Treatment Level Data. For each tattoo at the treatment level we detail four main parameters involved in laser-assisted tattoo removal, parameters that are actual tattoo-removal laser settings: fluence, spot size, wavelength and frequency. We briefly detail the treatment-level features in Table 3. Apart from observing complication occurrence/non-occurrence as a response variable in our analysis with respect to tattoo-level factors and patient demographic features, we also study the variation of treatment parameters (e.g., fluence, spot size) selected by practitioners given a particular tattoo characteristic such as tattoo color and age. The intention for viewing treatment parameters as a response characteristic is to gauge whether current clinician practices are at all particular to tattoo-level characteristics at a statistically significant level.

In our study, some additional design choices were made with regards to studying the treatment parameters over a given treatment sequence. Since the composition of tattoos have

Table 2
Table detailing tattoo level factors of interest.

Variable	Description
Category	Tattoo location on body (nominal)
Age	Age of tattoo (integer-valued)
Colors	Whether the tattoo was black/blue or had other colors (one-hot encoded as black-blue/not-black-blue, binary)
Professional	Whether the tattoo was done professionally, or by an amateur (one-hot encoded as professional/not-professional, binary)
Treatment Total	Total number of treatments recorded for a tattoo (integer-valued)
Fitzpatrick Score	Fitzpatrick skin tone rate: fairest tone at I to deeply pigmented at VI (ordinal)
Complications	Indication of whether the tattoo ever experienced any complication at all (one-hot encoded, binary)

great variation between and within different tattoo characteristics (e.g., color, ink composition and size) the number of treatments recorded varied greatly between tattoos. For example, a treatment sequence may have lasted over 10 weeks for one recorded tattoo, another may have had only a single appointment. Thus, two variations of variable transformation on the laser treatment parameters were made.

Table 3
Table detailing treatment level variables of interest.

Variable	Description
Fluence	Laser heat and energy intensity measured in joules/cm^2 (continuous)
Spot Size	Laser spot radius measured in millimeters (continuous)
Wavelength	Laser wavelength at two levels (532nm and 1064nm, binary)
Frequency	Laser frequency measured in Hertz at two levels (5Hz and 10Hz, binary)
Treatment Day	Days since first treatment (integer-valued)

The first variation is to simply record the mean and standard deviation of the above treatment settings for the entire period in which a tattoo had undergone treatment; this variation was used to study current clinician practices at Homeboy Industries to see how responses of laser settings changed (i.e., how clinicians selected different treatment parameters) on average as a result of varying tattoo characteristics. The second variation also records the means and standard deviations of the distribution of treatment parameters in sequence, but only up to the first instance of a complication occurring; naturally, if a complication never occurs in a tattoo, the mean and standard deviation of applied laser settings are computed over the entire treatment period. Figure 1 shows time series of fluence (in red), one of the four possible laser parameters, over different tattoos with data from the full time series up until the first-arrival of a complication—as well as change the of laser fluence between treatments (in blue), to be detailed below. Note the irregular duration of time intervals between treatments, as well as clear nonlinear fluctuations of fluence chosen by clinicians.

Furthermore, as an extension of the first complication arrival data, we create additional variables of means and standard deviations with forward finite differences. For a given variable, representing a time-series treatment parameter, we compute mean and standard deviations

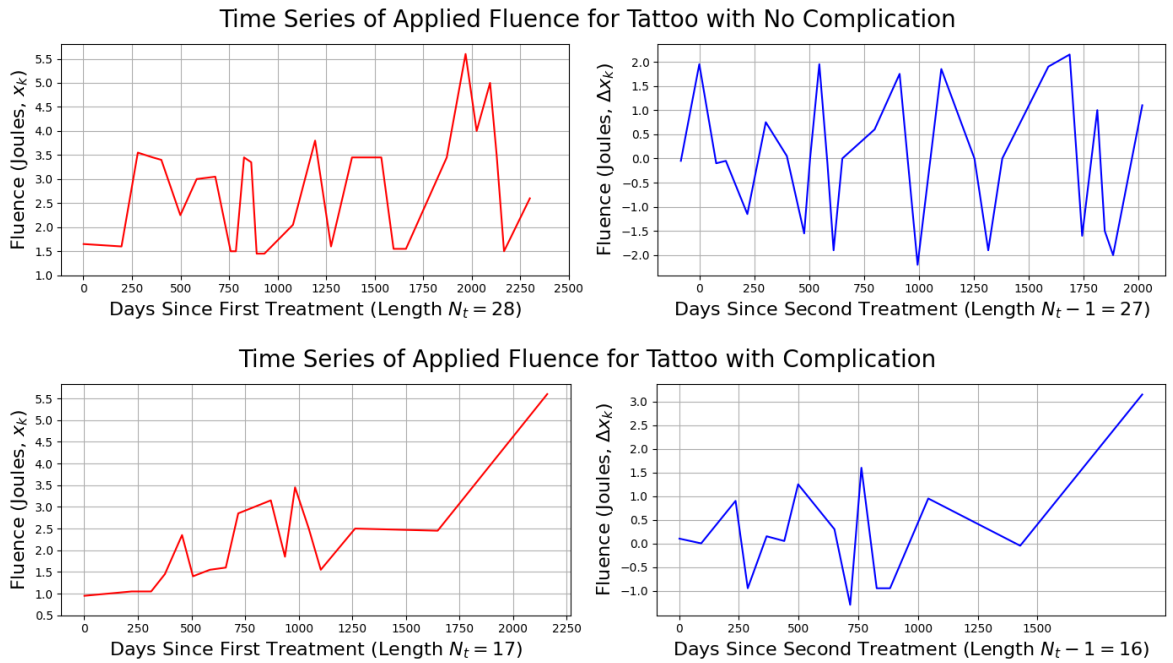

Figure 1. *We show fluence levels fluctuating by clinician application of laser-assisted tattoo removal. The plots in the first row are of fluence settings for a tattoo that does not develop any complications, while the plots in the second row are of one that has experienced complications. The red time series graphs represents the recorded fluence levels, while the blue time series graphs represent the corresponding forward differenced time series of fluence levels. Additionally, note that the time intervals are not uniform.*

for

$$\Delta x_k = x_k - x_{k-1},$$

where the terminating point in the time series, x_t, is dependent on whether we are observing first-arrivals of complications or the entire treatment sequence for a treated tattoo. For example, if we are observing laser fluence for first-arrivals data we produce a forward finite difference variable Δx, where the time series terminates at the index in the sequence corresponding to the treatment where a complication was first reported. Consequently, when computing the mean and standard deviations with respect to this forward finite difference our expressions simplify as

$$\overline{\Delta x} = \frac{1}{N_t - 1}\left(\sum_{k=2}^{N_t} \Delta x_k\right) = \frac{1}{N_t - 1}\sum_{k=2}^{N_t}(x_k - x_{k-1}) = \frac{1}{N_t - 1}\left(x_{N_t} - x_1\right),$$

where we are essentially measuring the average likelihood of a given treatment parameter setting from the first appointment to the terminal appointment (i.e., the full treatment sequence for original treatment-level dataset in its entirety, and the truncated treatment sequence for

the first-arrivals dataset). Similarly, for standard deviations of the differenced time series:

$$\hat{\sigma}_{\Delta x} = \sqrt{(S_{\Delta x})^2} = \sqrt{\frac{1}{N_t - 2}\sum_{k=2}^{N_t}\left(\Delta x_k - \overline{\Delta x}\right)^2} = \sqrt{\frac{\sum_{k=2}^{N_t}(\Delta x_k)2}{N_t - 2} - \left(\frac{N_t - 1}{N_t - 2}\right)\overline{\Delta x}^2},$$

where N_t is the number of treatments given to a particular tattoo (note the linear unbiased estimate of standard deviation). The purpose of this variable differencing is to eliminate any apparent trend; a first order forward differencing effectively applies a linear filter to our time series [22]. The plots of the time series in Figure 1 demonstrates the variation of x_k and Δx_k with respect to laser fluence. Tattoo removal clinicians generally aim to gradually increase the fluence over the treatment sequence, the effect of this difference is to essentially flatten this linearly trending process, and characterize the distribution of applied laser treatment settings into statistical parameters (e.g., mean and standard deviation).

2.4. Statistical Inference. We applied both parametric and nonparametric statistical tests of significance to identify statistically significant factors correlated to complication occurrence in our data (in all our tests we use $\alpha = 0.06$). We were interested in observing differences in proportions across different patient demographics and tattoo level factors. We applied the z-test for two proportion testing as well as the chi-square test of independence for more than two sample proportions. (It may seem unnecessary to use a z-test for proportions—and the data itself is from a "single" sample—, but we'd like to make the specification that patients that experienced a complication are distinct from those that did not, and that the z-test for proportions is consequently comparable to a chi-square test of factors with two levels.) The hypothesis statement generally for k-sample proportions is

$$H_0 : p_a - p_b = 0$$
$$H_a : p_a - p_b \neq 0$$

where p_a and p_b represent distinct sample proportions, pairwise, across k-samples. The necessary assumptions for these tests were found to be met before their application: independent samples and observations, sufficient sample sizes, and mutually exclusive categories in variables [17].

Along with these statistical descriptions of proportions, we additionally were interested in observing current distribution differences in removal practices and laser treatment settings. We use Welch's two-sample t-test to observe parametric differences of averages between different treatment settings by whether a tattoo experienced a complication or not [24, 25]. When the necessary assumptions for a t-test failed—observation independence, sufficient sample size and normality—we resorted to using the Wilcoxon rank sum test and the randomization test. We strictly applied the Wilcoxon rank sum test to responses that were categorical-ordinal on independent observations [16]. The null hypothesis for the Wilcoxon rank sum test is that the difference in central tendency of the distributions in our data is randomly occurring, with the alternative being that difference in central tendency is not random, at a statistically significant level.

In our implementation of the randomization test we were interested to see if the measured difference of sample averages for settings applied to tattoos that developed complications

versus those that did not were of random chance. Algorithm 2.1 details the randomization test.

Algorithm 2.1 Randomization Test

Let N be the number of times resampling occurs
Let $\Delta\mu := \mu_0 - \mu_1$ be the empirical difference of means between samples
for iteration $< N$ (sufficiently large) **do**
  Remove observation labels, creating null distribution, $H_0 \sim X$, in array
  Let X^*:=Permute(X), *randomly* permuting observations in array
  Split X^* by $\hat{p}_0$ and $\hat{p}_1$ label proportions for samples $X^*_{\hat{p}_0}$ and $X^*_{\hat{p}_1}$
  Let $\hat{\mu}^*_0$:=Mean($X^*_{\hat{p}_0}$) and $\hat{\mu}^*_1$:=Mean($X^*_{\hat{p}_1}$), randomized means
  Let (ΔX^*) be the randomized empirical distribution of differences
  Compute $\hat{\mu}^*_\Delta := \hat{\mu}^*_0 - \hat{\mu}^*_1$ and store into (ΔX^*)
end for
Let B be an array of true/false values corresponding to $\{(\Delta X^*) \leq -|\Delta\mu|, |\Delta\mu| \leq (\Delta X^*)\}$
Transform B into an array of numeric values 0/1; 1 for true, 0 for false
return Empirical p-value, $\hat{p}^*$:=Mean(B)

The null hypothesis of the randomization test is that the statistic of interest is produced by random chance, with the alternative being that the observed sample statistic is not randomly produced at a statistically significant level. Both the Wilcoxon rank sum and randomization tests allow for an interpretation of differences between factor levels that are correlated to complication rates, albeit with less statistical power being nonparametric with relative sample sizes between complication and noncomplication tattoos being heavily unbalanced, 2,000 to 118, respectively [9].

Finally, in comparing the means across multiple levels of a factor we applied one-way analysis of variance (ANOVA) where parametric assumptions for normal probablity distributions were satisfied, and a Kruskal-Wallis test for one-way ANOVA where these assumptions were violated. We verified the assumptions of normality, linearity and homoscedasticity [3, 5]. The hypotheses for one-way ANOVA are

$$H_0 : \mu_1 = \mu_2 = \cdots = \mu_k$$
$$H_a : \mu_i \neq \mu_j, \text{ for some } i \neq j$$

where the null hypothesis says that the population means between k groups of categories are equal, and alternative says that at a statistically significant level there exists a difference among the k population means. Applying ANOVA to our sample tattoo data we reject our null hypothesis, or do not reject our null hypothesis—in which case we then accept the alternative hypothesis. In order to locate the source of variation where shown to be significant, we applied Tukey's HSD test as a multiple comparison among the multi-level factors for which one-way ANOVA was applied [23]. (In our actual study below we prefer the Tukey-Kramer HSD over an alternative post hoc such as Fisher's LSD for multiple pairwise comparison in analyses involving a large number of factor levels (> 4) as the Tukey-Kramer HSD test generally controls the overall type 1 error rate—probability of falsely rejecting the null hypothesis

[18].) In particular, we used the Tukey-Kramer HSD test to account for unequal samples when computing the *studentized range distribution statistic* for the corresponding confidence interval

$$\bar{x}_{i\cdot} - \bar{x}_{j\cdot} \pm \frac{q_{\alpha;k;N-k}}{\sqrt{2}} \hat{\sigma}_\epsilon \sqrt{\frac{1}{n_i} + \frac{1}{n_j}}$$

where i and j represent distinct sample distributions from different populations; without loss of generality, $\bar{x}_{j\cdot}$ and n_j represents the sample average and sample size for the feature of interest, respectively. The coefficient $q_{\alpha;k;N-k}$ is the upper α percentage points of the studentized range statistic, q, defined as

$$q = \frac{\bar{x}_{\max} - \bar{x}_{\min}}{\sqrt{MS_E/n}}$$

where $\bar{x}_{\max}$ and $\bar{x}_{\min}$ are the largest and smallest sample means; N is the number of observations; k is the number of populations or levels in a factor; and $N - k$ is the degrees of freedom associated with the mean squared error, MS_E, in the studentized range statistic and n represents the number of observations [13]. Additionally, $\hat{\sigma}_\epsilon$ denotes the square root of the overall variation of the feature of interest in our total population.

When our parametric conditions failed (e.g., sample distribution normality and homoscedasticity), we resorted to applying the Kruskal-Wallis test for one-way nonparametric ANOVA [14]. The null hypothesis of this test is that the medians of our populations are equal, and the alternative is that the there is at least one source of random variation among our populations with a median that is statistically significant among some pair of populations (e.g., median fluence of settings applied to tattoos on the face versus tattoos on the neck or those on the upper extremeties). As a natural post hoc for multiple pairwise comparisons after the Kruskal-Wallis test we utilized Dunn's test. We find for the factors that failed the ANOVA assumptions, if the shape of continuous distributions between groups are nearly identical, we are permitted to interpret Dunn's test as a multiple pairwise comparison test of population median differences (i.e., test of statistically significant differences between population group medians based on provided sample data). We additionally apply the Benjamini-Hochberg correction for controlling the false discovery rate for this comparison, as opposed to a more strict p-value adjustment procedure such as Bonferroni's correction [2, 19].

2.5. Statistical Learning. In addition to performing inference, we were interested in developing statistical models of our data as related to complication occurrence (and nonoccurrence) across the first-arrivals tattoo-level dataset. In our data for complication occurrence, the actual response corresponds to a binary classification problem where complication occurrence in a given tattoo treatment sequence corresponds to a "true" outcome and complication nonoccurrence corresponds to a "false" outcome. Generally, logistic regression models a probabilistic relationship of the occurrence of some response along with its predictors [12]. More precisely, logistic regression models the following:

$$p(X_1, X_2, \ldots, X_n) = P(Y = 1 | X_1, X_2, \ldots, X_n),$$

where in our case we have a binary label response of complication one-hot encoded as "1" and complication nonoccurrence encoded as "0"; and there are $n = 10$ predictors, after heuristically

removing highly correlated variables using an absolute correlation threshold of 0.6. The p represents the probability that a given observation with its corresponding inputs are of an outcome relating to tattoo complication. The precise formulation of the logistic regression model is

$$p(X_1, X_2, \ldots, X_n) = \frac{e^{\beta_0+\beta_1 X_1+\cdots+\beta_n X_n}}{1 + e^{\beta_0+\beta_1 X_1+\cdots+\beta_n X_n}}$$

The coefficients of the logistic regression model are estimated using maximum likelihood estimation (MLE) [4]. These estimates asymptotically follow a normal distribution, from which we can infer statistical significance and evident dependence on the corresponding variables to the response in the data. We use this process to identify statistically significant predictors of treatment parameters corresponding to complications (again, with an $\alpha = 0.06$).

In addition to modelling our data to find statistically significant factors corresponding to complication occurrence, we wanted to produce a ranking of statistically significant variables in our data. We do this using decision trees, gradient boosting and bootstrapping [10, 6]. Decision trees are machine learning models that recursively stratify (split) a feature space—tattoo treatment parameters in our case—into regions that are estimated to be numerically associated to a given response variable. Boosting is a general statistical/machine learning paradigm that involves bootstrapping a dataset into many copies, and sequentially fitting multiple machine learning models to the dataset. In this boosting procedure on decision trees there are three tuning parameters: B, the number of trees; λ, a shrinkage parameter controlling the learning rate of the boosting method; and d, the number of splits in a decision tree, or variables to consider in the feature space stratification process—alternatively, d is also understood as the interaction depth, controlling the order of complexity of our decision trees (e.g., if $d = 1$, we have only a single layer of splits; if $d = 2$, we have a first layer of feature space stratification, followed by another; and so on).

Boosting with decision trees on a given dataset can produce ranks of variable importance through the mean decrease in Gini index, a measure of total variance across K classes, formulated as

$$G = \sum_{k=1}^{K} \hat{p}_{mk}(1 - \hat{p}_{mk}),$$

where $\hat{p}_{mk}$ represents the proportion of training observations in the mth stratified region corresponding to the labels of class k. The rank of variable importance is measured by the mean decrease in the Gini index over all the decision trees fitted in the boosting procedure, where the higher the mean decrease in the Gini index for that variable, the more important it is. From this procedure we additionally bootstrap our dataset by fitting these gradient boosted trees in our dataset 300 times, with each fit utilizing tuned hyperparameters from repeated k-fold cross validation.

k-fold cross-validation is a model hyperparameter tuning method that works by looking over a set of candidate tuning parameters and testing from all combinations of candidate settings to find the best hyperparameter configuration. More precisely, k partitions of the original dataset are made where over each partition, $\tilde{k}$, the model is fit over the complement $k-1$ partitions; the performance metric of choice—for our gradient boosted trees, we use accuracy for simplicity—is then averaged and recorded for a given hyperparameter combination; among

the sample of measurements of average performance, the hyperparameter configuration with the best average performance is selected for final model training and testing. Repeated k-fold cross validation does exactly all of the above, but extends the cross-validation process by "repeating" k-fold cross-validation on new permutations of data such that different data points are relocated to different partitions between k-fold processes. For gradient boosting, we tuned our model between 50, 100 and 150 trees; our shrinkage parameter was kept constant at 0.1; and the interaction depth was chosen between 1, 2 and 3.

Finally, we then produce an empirical distribution of ranks of variable importance for features considered in our model; the most frequent rank of a particular feature is then assigned as the feature's empirical importance rank. The final rankings of these features then decide which features tattoo removal practitioners should be most aware of in the tattoo removal treatment process.

3. Results.

We tabulate our results of statistical inference as well as present the results of our machine learning models. Throughout we discuss results and the particular application of our methods to data.

3.1. Statistical Descriptions.

Below are tables corresponding to two-sample (Table 4) and k-sample proportion tests (Table 5). We measure complication rates across different factors at the patient demographic and tattoo levels. In our tables, "Tattoo Complication" refers to whether a single tattoo ever experiences a complication (i.e., complication occurrence/non-occurrence) within a treatment sequence spanning multiple days.

Table 4

Details of two-sample proportions and p-values for z-test. Statistically significant factors determining significant proportion differences in complication rates differences are marked by the asterisk. The "True" column represents proportions that are characterized by the factor of interest; "False" column represents proportions that do not have the given factor characteristic.

Response Proportion *by* Factor	**True**	**False**	***p*-value**
Treatment Completion **by** Complication	0.14407	0.07300	0.0083*
Tattoo Complication **by** Colored Tattoo	0.10857	0.04996	0.0035*
Tattoo Complication **by** Professional	0.09351	0.04291	0.0010*
Complication **by** Sex (Male/Female)	0.09964	0.11764	0.6163
Complication **by** Patient Median Age	0.13061	0.08560	0.1381
Complication **by** Patient Ethnicity	0.12010	0.06723	0.1452
Complication **by** Tattoo Median Age	0.06379	0.08560	0.5562
Complication **by** Patient Fitzpatrick Score ($>$ III)	0.12000	0.10287	0.5562

In our application of Welch's t-test, Wilcoxon's rank sum and randomization tests, we observe the distribution differences of settings applied to the tattoos over treatments up until the first arrival of a complication in the treatment sequence; if a tattoo never experiences a complication the statistics of treatment parameters is computed across all treatments for that tattoo. Table 6 shows the results from our t-test as applied to distributions of time-series statistics (mean and standard deviation) computed over our settings that satisfies some heuristic observations of normality; Table 7 shows the results of Wilcoxon rank sum and randomization tests over distributions of sample statistics that did not satisfy normality and followed highly irregular distribution shapes (i.e., multimodal and skewed).

Table 5

Details of k-sample proportions and p-values for chi-squared tests of independence. Statistically significant factors determining significant proportion differences in complication rates are marked by the asterisk. We characterize each factor as four-level ordinal factors defined as quarters split by quartile values; we have a five-number summary as well, omitting trivial minimums, for each factor listed below from our sample.

Reponse *by* Factor	Parameter	1st Quarter	2nd Quarter	3rd Quarter	4th Quarter
Complication **by**	Summary	4 (Q1)	7 (Q2)	12 (Q3)	65 (Max)
Total # of Treatments	n_i	122	116	117	122
Tattoo (p=6.51e-10*)	$\hat{p}_i$	0.02417	0.03349	0.05181	0.12277
Complication **by** Total	Summary	4 (Q1)	7 (Q2)	12 (Q3)	100 (Max)
# of Tattoos per	n_i	92	70	87	100
Patient (p=0.08395)	$\hat{p}_i$	0.11957	0.18571	0.08046	0.07000
Complication **by** Total	Summary	11 (Q1)	24 (Q2)	52 (Q3)	72 (Max)
# of Treatments	n_i	122	116	117	122
Patient (p=1.35e-06*)	$\hat{p}_i$	0.04918	0.06034	0.09402	0.24590
Complication **by**	Summary	1.4375 (Q1)	1.9134 (Q2)	2.5445 (Q3)	4.5773 (Max)
Mean Fluence (J/cm^2)	n_i	508	505	509	509
(p=3.532e-05*)	$\hat{p}_i$	0.01771	0.04554	0.07466	0.07662
Tattoo Complication	Summary	3 yrs (Q1)	7 yrs (Q2)	12 yrs (Q3)	40 yrs (Max)
by Tattoo Age	n_i	292	306	233	253
(p=0.1316)	$\hat{p}_i$	0.03425	0.071895	0.05150	0.075099
Patient Complication	Summary	29 yrs (Q1)	34 yrs (Q2)	41 yrs (Q3)	80 yrs (Max)
by Patient Age	n_i	148	109	128	117
(p=0.2807)	$\hat{p}_i$	0.09459	0.07339	0.14844	0.11111

Table 6

We present the sample averages of these settings along with the corresponding p-values from Welch's t-test. We find no statistically significant treatment settings among these parameters.

Treatment Parameter	**Complication**	**No Complication**	***p*-value**
Mean Fluence	1.958739	2.013944	0.5713
Mean Differenced Fluence	0.1848074	0.1729800	0.8181
Mean Spot Size	4.875772	4.942978	0.4402
Mean Differenced Spot Size	-0.06056931	-0.08294852	0.6542

Additionally, we have our results for current clinician practices and approaches. Table 8 presents our results from the Kruskal-Wallis one-way ANOVA for treatment parameters with distributions that violated the assumptions required for parametric one-way ANOVA. We applied non-parametric one-way ANOVA to study current variations of treatment parameters that were are found to be of practical interest in the literature. In particular we were interested in how clinicians applied different settings based on the tattoo age quartile. We found that the average fluence, average spot size, and standard deviation of differenced laser frequency were all statistically significant treatment parameters from our nonparametric ANOVA. Applying Dunn's test for multiple pairwise comparisons with the Benjamini-Hochberg correction on these statistically significant treatment parameters—again, with hypotheses on significant differences between group medians—, we discovered statistical significance. For average fluence there was a statistically significant difference between tattoos from quartiles 1 and 4 ($p_{\text{adj.}} \approx 0.02136$); for average spot size no pairs were found to have a significant difference; and for standard deviation in differenced laser frequency, quartiles 1 and 2 ($p_{\text{adj.}} \approx 0.01610$),

Table 7

We observe nonparameteric differences of statistics of treatment settings that are irregularly distributed with respect to the "first-arrivals" of complications data. For the Wilcoxon rank sum test we compute the sample median, and for the randomization test we compute the sample average. For each treatment parameter we present the results in two rows for sample median ($\hat{\eta}$, Wilcoxon rank sum) and sample mean ($\hat{\mu}$, randomization). As an additional comment, we find that for some factors, although the medians were the same, the distribution shapes between groups were found to be relatively distinct enough—due in part to the large imbalance of complication (118 tattoos with complications), no complication labels (2000 tattoos without complications)—in our nonparametric tests to be able to acquire such small p-values.

Treatment Parameter	**Statistic**	**Complication**	**No Complication**	***p*-value**
Mean Wavelength	$\hat{\eta}$	1064	1064	9.197e-06*
	$\hat{\mu}$	1010.425	1050.169	0.0013*
Mean Differenced Frequency	$\hat{\eta}$	0	0	0.05684*
	$\hat{\mu}$	0.23336582	0.03502956	0.0312*
Spot Size Standard Deviation	$\hat{\eta}$	0.80475	0.72648	0.05307*
	$\hat{\mu}$	0.8165863	0.6899222	0.0056*
Mean Differenced Wavelength	$\hat{\eta}$	0	0	1.784e-06*
	$\hat{\mu}$	4.663210	1.571152	0.1722
Mean Frequency	$\hat{\eta}$	9.375	9.375	0.5299
	$\hat{\mu}$	9.056429	9.145180	0.2733
Average Days Between Treatments	$\hat{\eta}$	61	71.5	0.5928
	$\hat{\mu}$	95.78247	102.58342	0.3339

2 and 3 ($p_{\text{adj.}} \approx 0.03489$), 1 and 4 ($p_{\text{adj.}} \approx 0.00822$), and 3 and 4 ($p_{\text{adj.}} \approx 0.04335$) were found to have statistically significant differences between group medians.

Table 8

Sample medians by tattoo age quarters (i.e., tattoo age in first quarter, second quarter...) and Kruskal-Wallis p-values. All parameters listed were found to be statistically significant; in other words, clinicians exercised distinct application of laser parameters based on tattoo age. We include a five number summary on tattoo age as well.

Quarter	**1st**	**2nd**	**3rd**	**4th**	
Five Number Summary	0yrs (Min)	3yrs (Q1)	7yrs (Q2)	12yrs (Q3)	40yrs (Max)
n_i	292	306	233	253	***p*-value**
Mean Fluence ($\hat{\eta}_i$)	1.85	1.86	1.967708	2.15	0.001946*
Mean Spot Size ($\hat{\eta}_i$)	4.870833	4.833333	4.8	4.823	7.314e-05*
SD Differ. Frequency ($\hat{\eta}_i$)	2.5	2.236068	2.526044	2.236068	2.294e-05*

In our application of parametric one-way ANOVA for studying means, we found particular interest in the application of average fluence with regards to the location of the tattoo on the body. There indeed exists some statistically significant variation with how a clinician chooses a particular fluence level with respect to tattoo location. Across tattoo locations on the body, the distributions of fluence were found to be relatively constant in variation and normally distributed. We obtained a p-value of 3.61-e05 for our one-way ANOVA. Then, applying Tukey's HSD test among tattoo location, we found that tattoos applied to the upper extremities and face had the most significant variation in average fluence applied—with sample averages of 2.09J/cm^2 and 1.996J/cm^2, respectively—among other tattoo locations on the body such as lower extremities, back, chest, neck, head and abdomen.

3.2. Significant Factors. We report our machine learning results in this section. Table 9 shows the numeric results of rank from our bootstrapped gradient boosting decision trees as well as the coefficient estimates and corresponding p-values from our logistic regression model. It should be noted that for both the logistic regression and gradient-boosted decision tree models, we removed a few variables from consideration. The removed variables were found to have high absolute Pearson correlation coefficient values (> 0.6) with other variables in the first-arrival sample statistics dataset of treatment parameters. The intention of reducing the number of predictors was to simplify our model and reduce any potential inflation of explained variation in our results.

To elaborate on our variable filtering procedure, if k pairs of variables were found to be highly correlated, as defined just before, and there existed a common variable between the k pairs, only the common variable would be kept in the model (e.g., in correlated pairs (a, b) and (b, c), with correlation -0.9 and 0.65, respectively, variables a and c would be removed from the model). For unique highly correlated pairs of variables, such variables were eliminated based on actual feature content; for example, if a variable such as standard deviation spot size was found to be correlated with mean differenced frequency, and a variable characterizing laser frequency was completely absent from the current variable selection, we would prefer the mean difference frequency variable. The intention of keeping only the common variable within the pairs is to minimize model complexity by expressing the model with as few variables as heuristically possible. Additionally, we emphasize that the purpose of applying statistical machine learning was to essentially simulate clinician laser removal practices. Consequently, no patient demographic or tattoo characteristic features were considered as in clinical settings different clinicians would essentially meet with different patients randomly with overwhelming primary consideration of previous laser treatment history over factors such as patient age, skin tone and other patient demographic features. Furthermore, including demographic information effectively and entirely removed over half the data points due to missing data in our demographic features.

For our models we ultimately selected 10 predictors with the filtering method that we described; however, we separately built a distinct logistic regression model through purposeful selection and general variable selection procedures [26]. To be precise, we additionally considered patient demographic features such as Fitzpatrick score—even with the effect of omitting data points due to missing data—, and the total number of treatments on the tattoo, but these factors were all either insignificant in univariate logistic regression, or essentially created data leakage. For factors such as the total number of treatments on a tattoo, a patient may logically terminate the treatment sequence much earlier due to poor results compared to a patient that completes and goes through the full treatment sequence, thus creating data leakage. Yet, this process of deliberately building a logistic regression model produced a final model absent of interaction terms and feature transformations (e.g., polynomial splines, log and exponent transformations, etc...). Our deliberate process of model building included independently fitting each feature as univariate model; comparing different models using likelihood ratio tests (tests of similarity between models); diagnosing model residual deviance (model prediction error); and respecting the principal of parsimony, or preference over more economic models utilizing fewer features.

Our separate logistic regression model, distinct from the model developed with *a priori*

filtering, included the following features: average spot size ($p \approx 0.00010$), standard deviation in spot size ($p \approx 0.00549$), standard deviation in wavelength ($p \approx$ 1.41e-05), and standard deviation of days between appointments ($p \approx$ 9.11e-07). With a Hosmer and Lemeshow goodness-of-fit test—a common test for assessing similarities of a logistic regression model's performance compared to true labels—, we have a p-value of 0.08479 (> 0.06) indicating no statistically significant difference between observed and fitted values. Observing our model's ability to distinguish complications from no complication labels, we used the receiver operator characteristic (ROC) curve and the area under the curve (AUC), finding an AUC of approximately 0.731, where the AUC can be interpreted as the probability that given two data points of different labels, the input data of the "higher ranked" label (complication occurrence) will be correctly ranked higher than the lower ranked label (complication nonoccurrence) [7].

With regards to the variable importance ranking system, with over 300 simulations of fitting gradient boosted trees, we produced a discrete distribution of ranks for our variables, as detailed in Section 2.5. In cases where the mode of ranks were tied between two factors, these ties were broken by observing which rank mode was greater. For example, in Table 9, we found that mean wavelength and mean differenced wavelength had rank modes of 1; however, since the frequency of the rank mode for mean wavelength (100) was greater than the rank mode of the mean differenced wavelength (90), we reassign "total ranks" for mean wavelength as greater than the mean differenced wavelength. Figure 2 displays a few rank distributions of variable importance from our simulations.

Table 9

Results from boosted decision tree simulations, and coefficient estimates and p-values for logistic regression coefficients. Statistically significant logistic regression coefficients are asterisked. For our boosting procedure we show the ranking of important features corresponding to the mode of the variables' frequency distribution. Ties in mode frequency rank were broken by comparing relative frequencies. Statistically significant logistic regression coefficients are asterisked.

Treatment Parameter	Total Rank	Rank Mode	Estimate	p-value
Mean Wavelength	1	108 (1)	-0.004299	0.00118*
Mean Differenced Fluence	2	84 ($1 \to 2$)	0.550567	0.33951
Mean Differenced Wavelength	3	80 ($1 \to 2 \to 3$)	0.004979	0.35263
SD Spot Size	4	117 (4)	1.294140	0.00609*
Mean Days Between Appts.	5	93 (5)	-0.001966	0.35287
Mean Differenced Spot Size	6	76 ($5 \to 6$)	0.449331	0.38701
Mean Differenced Frequency	7	64 ($6 \to 7$)	0.466965	0.04217*
Mean Spot Size	8	75 ($7 \to 8$)	-0.154691	0.71596
Mean Fluence	9	106 (9)	-0.137015	0.59433
Mean Frequency	10	197 (10)	0.014290	0.93119

4. Discussion. For our tattoo data we applied a whole range of methods from classical statistical parametric tests to modern machine learning algorithms. We characterized the time series treatment data across all parameters using sample means and standard deviations, as well as having performed forward finite differencing to characterize the variation in laser-assisted tattoo removal treatment parameters *between* appointments. From the transformations made on our variables, we applied both parametric and nonparametric tests of

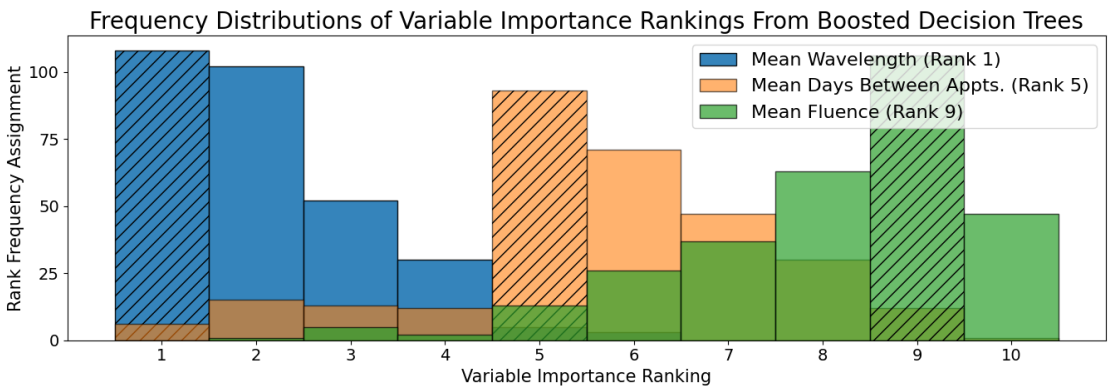

Figure 2. *Rank distributions of three variables with bars corresponding to "total rank" hatched. We assign the rankings for each treatment parameter based on the mode of their rank. We see that Mean Wavelength (Rank 1), Mean Days Between Appts. (Rank 5), and Mean Fluence (Rank 9) are of descending order of importance based on the location of the modes of ranks for each variable from 300 simulations.*

significance to identify tattoo-level and patient demographic characteristics strongly correlated to complication occcurence.

Below we list response-factor relationships that were found to be statistically significant along with a brief details with regards to practical implications:

- Complication occurrence is related to **decreased** likelihood of tattoo removal treatment completion. Patients may become ambivalent about continuing tattoo treatments if a complication is experienced.
- Tattoo removal complications are statistically **more likely** for tattoos done by a professional tattoo artist as opposed to an amateur artist. This may be due to the depth of the tattoo placement and/or ink utilized in the professional tattoo process.
- Tattoo removal complications are statistically **more likely** for tattoos with color (e.g., red, green, yellow) as opposed to black/blue tattoos. This may be due to the skin reaction sensitivity in tattoos with certain pigments.
- Greater total number of treatments on a tattoo and number of treatments a patient has undergone are all related to **increased** complication rates. With regards to tattoo age, clinicians apply statistically greater average fluences and have greater variability in spot size applied to tattoos over the treatment period.
- Greater average fluence applied to a tattoo is related to **increased** likelihood of complication occurrence.
- Greater average laser wavelength over a treatment period is *both* parametrically and non-parametrically related to **increased** likelihood of complication occurrence. Shorter wavelengths may be emulating intense radiation causing complications associated with skin discoloration.
- According to our logistic regression model, average wavelength, standard deviation in laser spot size, and overall change of laser frequency applied to a tattoo through a given

treatment sequence are related to **increased** likelihood of complication occurrence. It is also worth noting the response-factor relationships that were found to *not* be statistically significant with respect to the current literature (again, with brief explanations and implications):

- The average number of days between appointments is not statistically related to complication occurrence. The number of days between appointments in the data is sufficiently long for skin recovery before a follow-up appointment.
- Tattoo age and patient age is not statistically related to the variation in complication rates. This possibly suggests that current clinician approaches to tattoo removal take into account tattoo and patient age, since the application of certain settings used by clinicians vary between different ages at a statistically significant level.
- Between groups of tattoos that experienced complications and those that did not, the variation in average laser fluence, overall change in laser fluence between the first and last appointment, average spot size, and overall change in laser spot size between the first and last appointment were not found to be distinct at a statistically significant level.
- Patient sex, ethnicity, Fitzpatrick score, and being above the median age are not statistically related to complication occurrence within a recorded treatment sequence, independent of whether a patient completed the entire treatment process.

Finally, note the discrepancy of our results between the two machine learning models. Although our boosted decision tree procedure lowly ranks a statistically significant variable—according to our logistic regression model—such as "Mean Difference Laser Frequency," this result is not an inconsistent finding. The two models essentially offer distinct interpretations of the data.

The logistic regression model measures that in a total interaction of all ten variables, average wavelength, standard deviation in spot size, and average variation of laser frequency between the first appointment and pre-arrival of a recorded tattoo complication are statistically significant factors for the tattoo removal practitioner to be aware of in the treatment process. On the other hand, the aggregated boosted decision tree model ranks variables by observing how our predictions of complication occurrence/non-occurrence are affected if they are not considered in the interaction. For example, excluding average wavelength in the interaction of treatment parameters will most frequently negatively affect classification of complication occurrence/nonoccurrence in our data compared to the exclusion of average frequency which will be of least consequence in our predictions.

5. Conclusion.

Limitations worth noting with respect to our findings are that these results are specific to our given dataset, which has a fair share of missing values without any data imputation having been performed, and we do avoid making any full clinical interpretations of our results as well. Yet, our data is relatively abundant and detailed, and we've applied many interesting techniques along with applying this ad hoc aggregated variable ranking procedure using gradient boosted decision trees, which may prove to be a theoretically well-justified and statistically powerful way of producing ranks of variable importance.

Acknowledgements. RPY is grateful for the supervision, support and mentorship in this work from Professor Deanna Needell and Professor Jamie Haddock at the UCLA Computa-

tional Applied Mathematics group. The RPY also gives much thanks to Jessica Bogner at Homeboy Industries, Los Angeles, and Dr. Jo Marie Reilly, at Keck Medicine of USC, for providing data and giving guidance with this project. This work was partially supported by NSF BIGDATA #1740325, NSF DMS #2011140 and NSF DMS #2111440.

REFERENCES

[1] W. Bäumler and K. Weiss, *Laser assisted tattoo removal–state of the art and new developments*, Photochemical & Photobiological Sciences, 18 (2019), pp. 349–358.
[2] Y. Benjamini and Y. Hochberg, *Controlling the false discovery rate: a practical and powerful approach to multiple testing*, Journal of the Royal statistical society: series B (Methodological), 57 (1995), pp. 289–300.
[3] D. R. Cox, *Principles of statistical inference*, Cambridge university press, 2006.
[4] S. A. Czepiel, *Maximum likelihood estimation of logistic regression models: theory and implementation*, Available at czep. net/stat/mlelr. pdf, (2002), pp. 1825252548–1564645290.
[5] C. M. Douglas, *Design and analysis of experiments*, John Wiley and Sons Inc, 2001.
[6] B. Efron and R. J. Tibshirani, *An introduction to the bootstrap*, CRC press, 1994.
[7] T. Fawcett, *An introduction to roc analysis*, Pattern recognition letters, 27 (2006), pp. 861–874.
[8] T. B. Fitzpatrick, *Soleil et peau*, J Med Esthet, 2 (1975), pp. 33–34.
[9] P. Good, *Permutation tests: a practical guide to resampling methods for testing hypotheses*, Springer Science & Business Media, 2013.
[10] T. Hastie, R. Tibshirani, and J. Friedman, *The elements of statistical learning: data mining, inference, and prediction*, Springer Science & Business Media, 2009.
[11] S. G. Ho and C. L. Goh, *Laser tattoo removal: a clinical update*, Journal of cutaneous and aesthetic surgery, 8 (2015), p. 9.
[12] G. James, D. Witten, T. Hastie, and R. Tibshirani, *An introduction to statistical learning*, vol. 112, Springer, 2013.
[13] C. Y. Kramer, *Extension of multiple range tests to group means with unequal numbers of replications*, Biometrics, 12 (1956), pp. 307–310.
[14] W. H. Kruskal and W. A. Wallis, *Use of ranks in one-criterion variance analysis*, Journal of the American statistical Association, 47 (1952), pp. 583–621.
[15] I. Kurniadi, F. Tabri, A. Madjid, A. I. Anwar, and W. Widita, *Laser tattoo removal: Fundamental principles and practical approach*, Dermatologic Therapy, (2020), p. e14418.
[16] H. B. Mann and D. R. Whitney, *On a test of whether one of two random variables is stochastically larger than the other*, The annals of mathematical statistics, (1947), pp. 50–60.
[17] M. L. McHugh, *The chi-square test of independence*, Biochemia medica, 23 (2013), pp. 143–149.
[18] D. C. Montgomery, *Design and analysis of experiments*, John wiley & sons, 2017.
[19] T. V. Perneger, *What's wrong with bonferroni adjustments*, Bmj, 316 (1998), pp. 1236–1238.
[20] R Core Team, *R: A Language and Environment for Statistical Computing*, R Foundation for Statistical Computing, Vienna, Austria, 2017, https://www.R-project.org/.
[21] J. Serup and W. Bäumler, *Guide to treatment of tattoo complications and tattoo removal*, in Diagnosis and Therapy of Tattoo Complications, vol. 52, Karger Publishers, 2017, pp. 132–138.
[22] R. H. Shumway, D. S. Stoffer, and D. S. Stoffer, *Time series analysis and its applications*, vol. 3, Springer, 2000.
[23] J. W. Tukey, *Comparing individual means in the analysis of variance*, Biometrics, (1949), pp. 99–114.
[24] R. E. Walpole, R. H. Myers, S. L. Myers, and K. Ye, *Probability and statistics for engineers and scientists*, vol. 5, Macmillan New York, 1993.
[25] B. L. Welch, *On the comparison of several mean values: an alternative approach*, Biometrika, 38 (1951), pp. 330–336.
[26] Z. Zhang, *Model building strategy for logistic regression: purposeful selection*, Annals of translational medicine, 4 (2016).