Investigating the Efficacy of Anti-Androgens and Estrogens on Testosterone Suppression in Transgender Women

Rebecca M. Lindner, MS[1]
Columbia University, New York, NY, USA

Abstract

While cross-sex hormone therapy is often prescribed to transgender individuals, little research has been done into the efficacy of these drugs in achieving hormone levels consistent with gender identity. Transgender medicine is a relatively new field, and a number of barriers have often prevented individuals from receiving medical care, leading to few datasets. Through a large transgender primary care program, an anonymized cohort dataset is compiled of clinical information for several hundred transgender women to investigate whether predictive models can be created for change in testosterone level and testosterone suppression over the course of a hormone regimen, using a variety of linear and logistic regression techniques. Decision trees are also used to attempt to perform feature selection on the extracted data and determine the importance of various demographic and clinical indicators to testosterone levels.

1. Introduction

Transgender women often experience gender dysphoria in the gender assigned at birth, and are treated through some combination of hormone therapy and surgical procedures. Hormone therapy is a necessary precursor to surgical interventions, which for transgender women entails testosterone suppression and estrogen increases in order to feminize physical and psychological traits. The number of transgender women in the US is not known precisely, but studies put it at between 5 and 500 per 100,000 persons[1], meaning that more research into the effectiveness of the combined anti-androgen and estrogen therapy usually prescribed is critical to care. A number of studies have investigated harmful side effects of these off-label uses of these drugs[3], and some research has been done into the variety of treatment options available[2]. A few small studies have looked at testosterone levels in transgender women; however, standards of care have not been well established[1], and a variety of anti-androgens (spironolactone, finasteride, bicalutamide, leuprolide) and estrogens (medroxyprogesterone, micronized progesterone, estradiol, estradiol valerate) are in use.

The most quantifiable indicator of feminization due to hormone therapy is testosterone level. The normal testosterone ranges for cisgender men and women are 200-300 ng/dL and <75 ng/dL[1], respectively. Clinicians consider a "suppressed" testosterone level for transgender women to be <100 ng/dL. While studies estimate that the highest level of feminization occurs approximately 2-5 years into treatment[4], only a year of continuous hormone therapy supervised by a physician is required prior to pursuing surgery. Many patients (up to 50%) self-medicate with "street" hormones prior to seeking professional care[1], making it difficult to determine baseline testosterone levels. Due to this and other stigmatizing factors inherent in transgender medicine, there has been a paucity of datasets and research in this area.

This paper investigates whether clinical data can be used to predict raw and binarized testosterone level at the end of a hormone therapy regimen, as well as the raw and binarized change in testosterone level. There are two hypotheses: (1) testosterone levels in transgender women are associated with a small number of clinical indicators, and (2) the most important factors to testosterone suppression/decrease in testosterone are time on hormone therapy, time on particular regimen, initial testosterone value, and anti-androgen drugs.

2. Materials and Methods

2.1 Dataset Extraction and Construction

Initial datasets were pulled for feminizing hormone therapy regimens, looking at 8 different drugs at 22 different dosages for a total of approximately 224,000 records, and testosterone tests, a total of approximately 117,000 records. The intersection of these two sets yielded 655 unique patients with ICD-9 codes identifying them as transgender in their problem list and/or medical history.

A number of the 655 patients were excluded due to history of orchiectomy, diagnoses of prostate cancer (as many anti-androgen drugs studied here are also used as a treatment for prostate cancer), age below 13, or erroneous documentation of the individual as transgender. Due to these exclusions, as well as issues with missing data and lack of baseline testosterone levels, only 230 patients were selected for analysis. These 230 patients had 259 unique anti-androgen/estrogen regimens prescribed to them for at least three months, with baseline and final testosterone levels present. Each patient was followed on one (n=218), two (n=39), or three (n=2) unique regimen(s), and each regimen contained up to three different anti-androgen and estrogen drug formulations. The most commonly prescribed drugs were spironolactone, an anti-androgen, and estradiol, and estrogen replacement (Table 1).

2.2 Dataset Features

From the data extracted on regimens and testosterone levels, four variables were created in addition to binary variables for each of the drugs and their dosages (i.e. Finasteride, Finasteride_1, Finasteride_5): time on hormone therapy (in days), time on particular regimen (in days), initial testosterone level for the regimen, and a binary variable marking whether the regimen was the first for the patient. Additional data was extracted for the selected patients, including race and ethnicity, BMI, age, comorbid conditions, and other medications. Race and ethnicity were separated into binary variables for Hispanic/Latino(a), Black/African American, White, and Other Race, using only these categories due to missing information for many patients. BMI was coded both as a continuous variable and four binary variables for Underweight, Normal Weight, Overweight, and Obese, though only the continuous or binary variables were used in any one model. Age was coded only as a continuous variable. Select comorbid conditions were encoded as binary variables for HIV/AIDS, Cancer, Hepatitis C, Diabetes, Mental Illness, and Substance Abuse Disorders. The most common co-occurring medications were also encoded as binary variables, including ARVs, Statins, Antibiotics, Warfarin, Insulin, Nicotine Replacement, and Multivitamins. A total of 57 features were included in the final dataset, 52 of which were binary or binary encodings of categorical variables, and 5 of which were continuous.

Table 1. Anti-androgen and estrogen drugs prescribed to transgender women, with total regimens containing drug and dosages.

Category	Drug	Total Regimens	Dosage 1	Dosage 2	Dosage 3	Dosage 4	Dosage 5
Anti-Androgen	Spironolactone	172	25 mg	50 mg	100 mg		
Anti-Androgen	Finasteride	54	1 mg	5 mg			
Anti-Androgen	Leuprolide	31	7.5 mg/mL	11.25 mg/mL	15 mg/mL	22.5 mg/mL	30 mg/mL
Anti-Androgen	Bicalutamide	5	50 mg				
Estrogen	Estradiol	110	0.1 mg	0.5 mg	1 mg	2 mg	
Estrogen	Estradiol Valerate	43	10 mg/mL	20 mg/mL	40 mg/mL		
Estrogen	Micronized Progesterone	10	150 mg				
Estrogen	Medroxyprogesterone	5	5 mg	10 mg	150 mg		

Table 2. Contingency table of patients in dataset with suppressed/unsuppressed testosterone at final level and decrease/increase in testosterone.

	Testosterone Suppressed	Testosterone Unsuppressed
Testosterone Decrease	94	66
Testosterone Increase	19	80

2.2 Dataset Targets

The final dataset contained four target variables, two binary (testosterone suppressed at final value in regimen, and testosterone decreased from initial to final value) (Table 2) and two continuous (final testosterone value, and change from initial to final value). Initial testosterone values ranged from 5 to 3,372 ng/dL, and final testosterone values from 5 to 3,678 ng/dL. Changes in testosterone ranged from -3,080 ng/dL to +2,897 ng/dL.

2.2 Methods

Using a variety of regression methods, including logistic regression with L1 and L2 penalties, least squares linear regression, LASSO regression, ridge regression, and elastic net regression, predictive models were created from the dataset for both the binary and continuous target variables. The most successful models, along with a decision tree classifier, were then used to identify the variables most important to predicting a decrease in testosterone and/or a suppressed testosterone level and perform feature selection for future research.

3. Results

3.1 Linear Regression

Each linear regression method was performed both on unstandardized data and on data with the features standardized by subtracting the mean of each and dividing by the standard deviation, and target variables centered by subtracting the mean. This type of standardization performed best across multiple tests in each regression style. The R^2 score was computed for each regression by randomizing the data into train and test sets with an 80/20 split, and running the model 100 times to determine mean and confidence intervals for each score (Table 3). Validation was repeated for 100 iterations in order to stabilize the predictions of the less accurate models, such as Least Squares and Ridge.

Table 3. Mean R^2 scores computed for each linear model through randomized holdout validation repeated 100 times.

	Least Squares		**LASSO**		**Ridge**		**Elastic Net**	
	Unstandardized	Standardized	Unstandardized	Standardized	Unstandardized	Standardized	Unstandardized	Standardized
Change in T Level	0.113 *95% CI:* *(0.056, 0.170)*	0.00 *95% CI:* *(0.00, 0.00)*	0.229 *95% CI:* *(0.185, 0.272)*	0.159 *95% CI:* *(0.108, 0.209)*	0.211 *95% CI:* *(0.168, 0.255)*	0.126 *95% CI:* *(0.071, 0.181)*	0.288 *95% CI:* *(0.252, 0.324)*	0.337 *95% CI:* *(0.309, 0.365)*
Final T Level	-0.094 *95% CI:* *(-0.221, 0.033)*	0.000 *95% CI:* *(0.00, 0.00)*	0.035 *95% CI:* *(-0.084, 0.153)*	-0.039 *95% CI:* *(-0.160,0.082)*	0.013 *95% CI:* *(-0.107, 0.134)*	-0.075 *95% CI:* *(-0.198,0.049)*	0.150 *95% CI:* *(0.073, 0.227)*	0.241 *95% CI:* *(0.207, 0.274)*

Table 4. Mean accuracy scores computed for each logistic model through 10-fold cross validation.

	L1 Penalty		L2 Penalty	
	Unstandardized	Standardized	Unstandardized	Standardized
Change in T Level	0.707 *95% CI:* *(0.642, 0.771)*	0.699 *95% CI:* *(0.634, 0.764)*	0.687 *95% CI:* *(0.627, 0.747)*	0.687 *95% CI:* *(0.627, 0.747)*
Final T Level	0.710 *95% CI:* *(0.657, 0.763)*	0.690 *95% CI:* *(0.639, 0.742)*	0.722 *95% CI:* *(0.681, 0.763)*	0.702 *95% CI:* *(0.647, 0.757)*

3.2 Logistic Regression

Logistic regression was performed on the two binary target variables – testosterone suppressed (1 if final testosterone value <100 ng/dL, 0 otherwise), and testosterone decreased (1 if final testosterone was less than initial testosterone, 0 otherwise). Both L1 and L2 penalties were used on unstandardized and standardized datasets, where continuous features were standardized again by subtracting the mean and dividing by the standard deviation. Binary features were left unstandardized in both iterations. Evaluation was performed through randomized 10-fold cross validation, and accuracy and precision scores remained fairly stable (Table 4, 5).

Table 5. Mean precision scores computed for each logistic model through 10-fold cross validation.

	L1 Penalty		L2 Penalty	
	Unstandardized	Standardized	Unstandardized	Standardized
Change in T Level	0.735 *95% CI:* *(0.693, 0.774)*	0.728 *95% CI:* *(0.689, 0.768)*	0.719 *95% CI:* *(0.678, 0.760)*	0.718 *95% CI:* *(0.675, 0.760)*
Final T Level	0.687 *95% CI:* *(0.626, 0.747)*	0.657 *95% CI:* *(0.590, 0.725)*	0.691 *95% CI:* *(0.643, 0.739)*	0.667 *95% CI:* *(0.599, 0.734)*

3.3 Decision Trees

For the decision tree algorithm, the features were simplified to remove the dosages of each drug as well as the BMI binarized categories, leaving 29 features for use in predicting the two binary targets, testosterone suppression and decrease in testosterone. The decision trees were validated using randomized 10-fold cross validation. For a decrease in testosterone, a decision tree at a maximum depth of four had an accuracy of 0.606 (95% CI: [0.538, 0.675]) and a precision of 0.682 (95% CI: [0.626, 0.737]). For testosterone suppression, a decision tree again pruned to a maximum depth of four had an accuracy of 0.696 (95% CI: [0.654, 0.737]), and a precision of 0.675 (95% CI: [0.611, 0.739]). These trees are shown in Figures 1 and 2.

3.4 Feature Selection

Feature selection was performed using the best performing predictors, as well as the decision trees. For change in testosterone level, these included logistic regression with an L1 penalty, and elastic net regression. For testosterone suppression, these included logistic regression with an L2 penalty, and elastic net regression. The coefficients for the best-performing version of the features (standardized or unstandardized) were tabulated, and the ten most influential features in each model were selected as important in each (Tables 6-8).

Figure 1. Decision tree for a decrease in testosterone, pruned to a depth of 4.

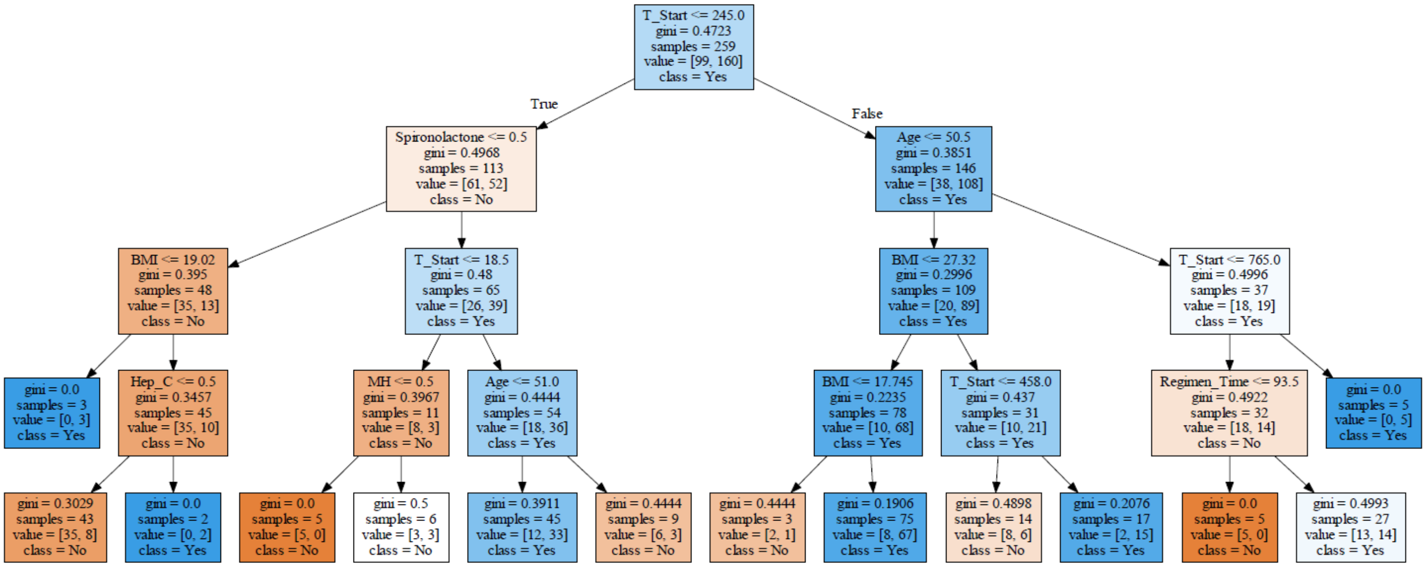

Figure 2. Decision tree for testosterone suppression (<100 ng/dL), pruned to a depth of 4.

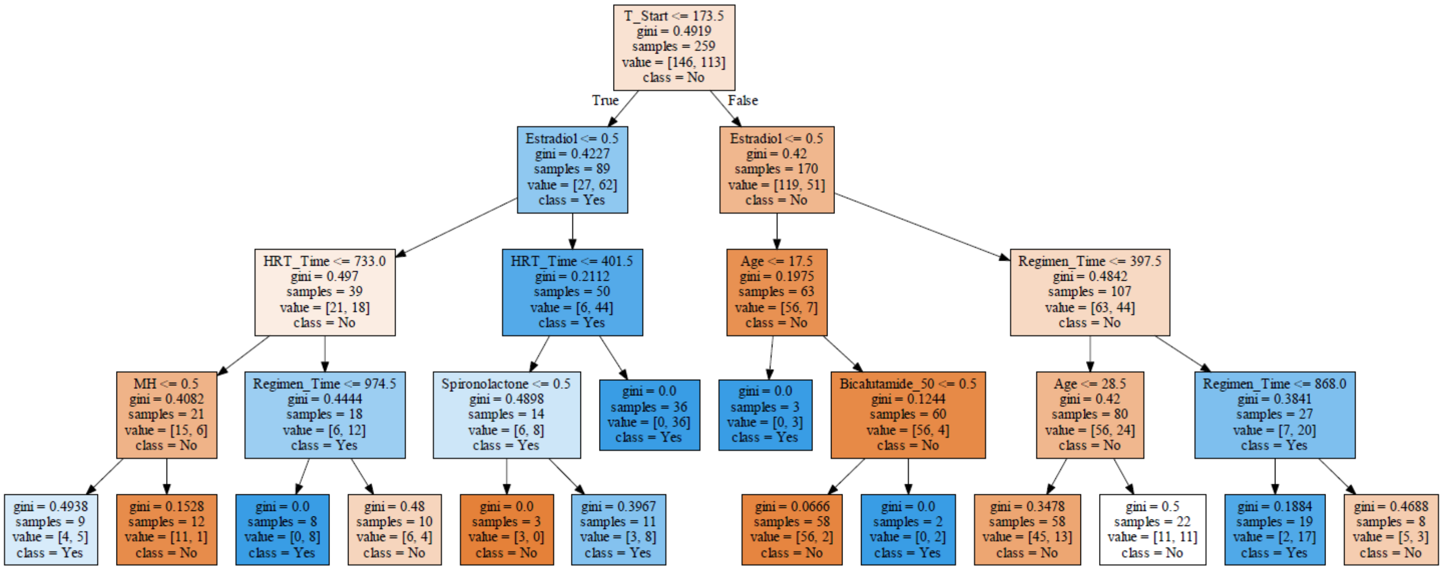

Table 6. 10 largest coefficients for standardized elastic net regression for change in testosterone value.

Feature	Coefficient
Initial Testosterone	131.937
Statin	-46.182
Age	-37.398
Hepatitis C	37.242
Estradiol	31.304
Cancer	29.579
Multivitamin	22.353
Nicotine	-22.29
Time on Regimen	-20.840
Finasteride (5)	-19.135

Table 7. 10 largest coefficients for unstandardized logistic L1 regression for change in testosterone value.

Feature	Coefficient
Medroxyprogest.	-1.925
Hepatitis C	1.375
Multivitamin	1.020
Cancer	0.807
Statin	-0.783
Spironolactone(100)	0.645
Estradiol	0.631
White	0.582
Obese	0.422
First Regimen	0.399

Table 8. 10 largest coefficients for standardized elastic net regression for testosterone suppression at end of regimen.

Feature	Coefficient
Initial Testosterone	77.757
Statin	40.816
Age	37.237
Estradiol	-36.885
Cancer	-31.086
Time on Hormones	-26.074
Medroxyprogest.(10)	25.455
Hepatitis C	-23.968
Estradiol Valerate(2)	-22.657
Finasteride (5)	20.543

Table 9. 10 largest coefficients for unstandardized logistic L2 regression for testosterone suppression at end of regimen.

Feature	Coefficient
Multivitamin	1.103
Hepatitis C	1.088
Medroxyprogest.	-1.086
Statin	-0.869
Cancer	0.755
First Regimen	0.722
Estradiol(1)	0.692
Spironolactone	0.654
White	0.633
Spironolactone(100)	0.615

4. Discussion

This investigation found that it is difficult to predict raw testosterone levels and raw change in testosterone from a small dataset with a limited number of features. Elastic net regression performed best with the data, combining the L1 and L2 penalties, potentially due to its ability to handle the collinearity brought about by the drug encodings. However, the R^2 values for this method were still well below 0.5 (Table 1), signaling that a more refined model with a larger dataset is needed. Change in testosterone value functioned slightly better as a target parameter in all linear regression models, showing that the variation in natural testosterone levels makes their raw values difficult to predict accurately. In all models except for the elastic net, the confidence interval for the R^2 for raw testosterone level included 0, meaning that none of the variation was explained by any of the selected features.

Logistic regression models showed more success with binarized target variables, with similar results for testosterone suppression and decrease in testosterone, in contrast to the linear models. This is likely due in part to the fact that the binarization removed much of the variability in natural and baseline testosterone levels by simplifying the variables. The best accuracy result was obtained for logistic regression with an L2 penalty run on unstandardized data (0.722, 95% CI: [0.681, 0.763]). Past studies have been performed on predicting hypogonadism, a similar research question, one of which showed an accuracy of 0.845[5], significantly higher than these models. However, these results are promising and indicate that logistic regression with some modifications may be able to predict testosterone levels in clinical settings. Since the L2 and L1 penalties had similar results, this may show that a number of interrelated features are at play, but predictions can be made with only some of the variables equally accurately.

Decision trees were markedly less successful than logistic regression models, and thus might provide insight only in terms of feature selection. Interestingly, the trees showed different decision points for predicting decrease in testosterone and testosterone suppression. The initial branch for change in testosterone was at starting testosterone level, splitting at 245 ng/dL. As this is mid-range for cisgender men, this may show that transgender women with naturally high testosterone are more likely have poor responses to anti-androgen medications. The rest of the tree branched at spironolactone, age, BMI, time on regimen, mental health disorders, and hepatitis C. While there are no known links between these variables (with the exception of spironolactone, an anti-androgen, and age) and testosterone levels, it is possible that there is a link or a confounding effect. The initial branch for testosterone suppression was again at starting testosterone level, but at 173.5 ng/dL. However, the next split was at estradiol, which should not have an effect on testosterone levels. Therefore it appears that testosterone level is difficult to predict logically through a decision tree model, though the model did perform slightly better for this target variable.

The features selected as most important by the two best performing models in each category were more consistent for change in testosterone, with five features shared between elastic net and logistic L1 regression (estradiol, cancer, statins, hepatitis C, and multivitamins). Multivitamins were likely a confounding variable, as patients taking multivitamins are more adherent to regimens in general. Fewer features were consistent for testosterone suppression (statins, cancer, and hepatitis C). These were also the only features shared by all four of the most accurate models, which seems to indicate that the variables known to be relevant (initial testosterone level, hormone therapy medications prescribed, and length of time on regimens) may have been confounded by the extra demographic variables, and the further research is necessary to understand potential connections.

4.1 Limitations and Future Work

This research was severely limited by the small size of the dataset and the lack of previous studies in the area. Most of these results were fairly inconclusive, primarily due to the wide variation in testosterone levels both before and after hormone therapy. Future work will be performed attempting to create models using the features selected by the best performing models here, potentially excluding outliers in order to better standardize the dataset.

5. Conclusion

The models tested (least squares, LASSO, ridge, and elastic net regression, logistic regression with L1 and L2 penalties, and decision trees) showed poor to fair results in predicting testosterone suppression and decrease in testosterone for hormone regimens prescribed to transgender women. The hypothesized features of importance (hormone therapy, time on particular regimen, initial testosterone value, and anti-androgen drugs) were not found to have a conclusive effect, while several other variables not known to be related to testosterone levels were found significant (statins, cancer, and hepatitis C). More research is required on the diverse population of transgender women in order to understand the effects of various features on testosterone levels and provide a clinically relevant model.

References

1. Tangpricha, V. & den Heijer, M. Oestrogen and anti-androgen therapy for transgender women. *Lancet Diabetes Endocrin.* 5: 291-300 (2017).
2. Mamoojee, Y., Seal, L. J., & Quinton, R. Transgender hormone therapy: understanding international variation in practice. *Lancet Diabetes Endocrin.* 5: 243-245 (2017).
3. Weinand, J. D. & Safer, J. D. Hormone therapy in transgender adults is safe with provider supervision; A review of hormone therapy sequelae for transgender individuals. *J. Clin. & Trans. Endocrin.* 2: 55-60 (2015).
4. Wesp, L. M. & Deutsch, M. B. Hormonal and Surgical Treatment Options for Transgender Women and Transfeminine Spectrum Persons. *Psychiatr. Clin. N. Am.* 40: 99-111 (2017).
5. Lu, Ti et al. Applying Machine Learning Techniques to the Identification of Late-Onset Hypogonadism in Elderly Men. *SpringerPlus* 5.1: 729 (2016).