TWO BAYESIAN APPROACHES THAT MODEL LONGITUDINAL ENDOGENOUS STEROID PATTERNS AND ABNORMAL CORRECTED T/E VALUES OF NFL PLAYERS

By

Dylan Burton Paulsen

A project submitted to the faculty of The University of Utah in partial fulfillment of the requirements for the degree of

Masters of Statistics, Econometrics

Department of Economics

The University of Utah

June 2010

TABLE OF CONTENTS

ABSTRACT

The detection of synthetic anabolic steroid doping continues to be a challenge for laboratories and anti-doping organizations alike. Profiling steroid use for individual athletes is rapidly becoming a valuable tool in this endeavor. Population based studies have been used by the World Anti-Doping Agency (WADA) to establish critical limits for various steroid concentrations and ratios. In addition, individual profiling is increasingly being utilized as a more precise and consistent tool for monitoring single athletes and catching errant behavior in smaller populations. The objective of this research is to use retrospective collected steroid data from the National Football League (NFL) and develop an adaptive Bayesian model (or known-sigma model) that will be used to evaluate longitudinal steroid profiles of individual athletes. In addition, a second model (the Gibbs sampling algorithm) will also be developed utilizing a Monte Carlo Markov Chain algorithm as a comparator based on the premise that there still exists large biological variability even in the NFL population. The database consists of over 17,000 samples from nearly 3,100 individual players. In addition, there are more than 600 players with 10 or more tests, which provide a large sample set suitable for individual longitudinal profiling. The two Bayesian approaches model corrected testosterone/epitestosterone concentration patterns, which is a key indicator of synthetic anabolic steroid abuse. Two athletes are modeled for this study: one athlete that has always tested

negative and one athlete that had a positive test occurrence. Both models are successful profiling each individual, but the Gibbs sampling procedure is the most accurate model since it treats the first initial tests of an athlete with very little precision and certainty, which is expected given the unknown physiological variations within the NFL population.

LIST OF TABLES

Table

LIST OF FIGURES

Figures

ACKNOWLEDGEMENTS

Funding for this research has been provided by the Partnership for Clean Competition (PCC). Also, data have been provided by the Sports Medicine Research and Testing Laboratory (SMRTL), which is a World Anti-Doping Agency (WADA), accredited laboratory. Finally, the National Football League and its associated players have agreed to allow their samples to be modeled. I also want to thank Richard Fowles, PhD for his help with the project.

CHAPTER 1

INTRODUCTION

Introduction

Enhancing one's performance and giving oneself a competitive advantage is as old as evolution itself. Therefore, athletes that introduce artificial substances into their systems in order to give themselves a competitive advantage are very much a common problem in today's competitive sports. For years health professionals have noted the physical and mental long-term consequences of using artificial steroid inducements to enhance athletic performance, yet steroid-doping still persists in both amateur and professional athletics alike. The consumer base demands statistical performance, and steroids represent a productive input that can enhance performance at relatively little cost. Therefore, the motivation for steroid abuse is prevalent and the consequences seem minimal if the player is caught. Many athletic organizations have mandated steroid testing for all players within their respective organizations in order to level the playing field for all athletes and base performance strictly on skill and hard work. Recent empirical studies suggest that punishment in some form deters negative behavior even if market forces only reward output.[1] In the case of professional athletes, the ramifications if caught doping can range from censure to even expulsion and the financial impact can be great since

endorsements are usually lost and public-presence is minimized. In any case, athletes will continuously adopt newer methods that enhance athletic performance and are undetectable using standard steroid-doping test protocols.

One subset of compounds used for performance enhancement is synthetic anabolic steroids, which metabolically mimic endogenous anabolic steroids that are naturally produced in the human body. By their very nature, synthetic anabolic steroids are difficult to identify in an analytical test setting due to three factors: first, they show up in very small quantities, usually in the range of ng/mL; second, synthetic anabolic steroids mimic the structural and chemical characteristic of endogenous anabolic steroids which are byproducts of natural biochemical pathways within the human body and can vary widely based on the physical and genetic characteristics of the tested individual; and, finally, human characteristics such as weight, age, general health, genetics, and even the time of day all play a role with metabolic buildup and breakdown of various endogenous anabolic steroids, especially testosterone[2]. The difficulty with such testing is to take a general rule and apply it to a large group of individuals while securing uniformity and fairness within the mandated requirements advocated by specific sporting organizations and anti-doping agencies alike; all, of which, require censoring of athletes who are caught doping. Sota et al has built a preliminary statistical model under controlled conditions in order to highlight testosterone doping, focusing on corrected testosterone to

epitestosterone ratios.[3] His test population consisted of northern European men under controlled conditions in which certain individuals were given testosterone directly in order to monitor their longitudinal T/E ratios. The objective of his research, as well as the research in which this current paper is based, is to build a model using a Bayesian approach that is unique to the athlete from a specific population and will flag a positive corrected T/E test based on the historical statistical trends for that specific athlete. There are two approaches which will be taken: one involves using the population characteristics of the athlete, thus building an adaptive model based on a known-sigma of an athlete's respective population; the other model is to utilize a Markov-Chain Monte-Carlo algorithm (MCMC), called the Gibbs sampler, and generate a parameterized model. In each case, the approach of this research is to take an individual and build and adaptive model that will show steroid abuse for that specific individual in a longitudinal manner based on new information as it becomes available.

The population under study for this model is composed strictly of players of various ages, races, and other physical characteristics that represent the National Football League (NFL). The sample size covers half the league's overall US population tested from January 1, 2006 to December 31, 2008. Most players were tested multiple times over multiple years. In summary, there are 8,634 tests for 637 NFL players that have had ten or more tests during the given time interval. Table 1. Summarizes the number of tests conducted from January 1, 2006 to December 31, 2008.

Table 1: Database Characteristics of the NFL Population

	Number of Samples	Number of Players/Year
Jan 1 – December 31, 2006	4,418	1,544
Jan 1 – December 31, 2007	6,862	1,779
Jan 1 – December 31, 2008	6,419	2,028
Total	17,699	3,108

The database characteristics for the NFL population tested from January 1, 2006 to December 31, 2008. This table represents the number of players available; it does not denote, however, the number of players that have every specific compound available to model.

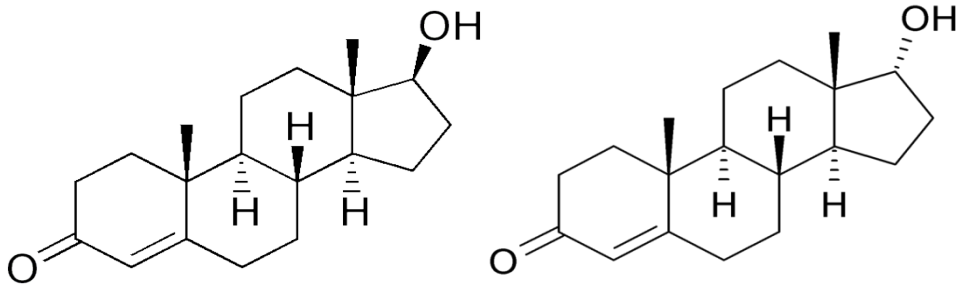
Even though there are nine characteristic compounds and ratios that can be examined, this model will initially focus on testosterone and epitestosterone levels in the form of a ratio called corrected T/E.

Testosterone is a naturally occurring endogenous anabolic compound found in both males and females. Although males naturally create on average ten times the concentration of testosterone versus their female counterparts, the compound does offer biological advantages to both genders. The advantages of using testosterone can include increased muscle density, linear body growth, increased protein storage, increased endurance, and increased bone density[4]. There is also a general belief that testosterone increases aggression and competitive attributes within an individual, even though there is still a high level of scientific debate directly related to whether these

two characteristics can be modified much by testosterone concentrations over short periods of time. Nonetheless, there is still a competitive advantage for an athlete to increase his/her testosterone exposure, so there is still a natural tendency for athletes within competitive sports to artificially increase testosterone levels from supplements that can easily be obtained on the open market. Another compound that is naturally found within the human body is epitestosterone. The biological pathway in which this compound is created is still under scientific investigation; however, the fact that testosterone and epitestosterone are diastereomers that differ in configuration of one stereogenic center (the 17β-hydroxyl group in this case) might indicate that that epitestosterone is a mirror-like byproduct from the same biological processes that generates testosterone (see figure 1). Although epitestosterone is the inactive epimer of testosterone, it is still banned by anti-doping agencies as a masking agent. In most normal human males, testosterone and epitestosterone exist as a 1:1 ratio, although levels much higher have been known to naturally occur in some non-doping cases. In fact, one NFL test subject for this paper had a ratio as high as 11.69 and was still confirmed negative by IRMS spectroscopy. The World Anti-Doping Agency (WADA) has determined any corrected T/E level above four constitutes possible testosterone doping and the result of any athlete caught and confirmed with levels that high without any medical justification can mean suspension for up to two years.[5] The detection of actual synthetic endogenous anabolic steroid peaks from mass spectra output can easily be

missed from the exorbitant amount of other steroid data that come with it; in the case of a typical WADA scenario, there can be up to 95 steroid compounds and their metabolites that come from one spectra. If an athlete is tested positive and confirmed as such, then the result can be the loss of sponsors, reputation, and even the ability to compete for an extended number of years beyond two (for example, the Olympics cycle every four years). It is imperative that testing is accurate for both the reputation of the lab as well as fairly treating athletes that are not tested positive and those that may have erroneously been tested positive (although there are plenty of internal controls that prevent this, but nothing is guaranteed).

There are many stratagems than an athlete can employ in order to mask a positive test when he/she willingly dopes. One such method is to mask testosterone doping by adding epitestosterone in order to dilute a high ratio; however, there are IRMS methods used to examine this approach as well.



Testosterone Epitestosterone

Figure 1: The molecular structures of testosterone and epitestosterone. The only difference between these two molecules is the angle of the 17β-hydroxyl group. Testosterone actively bonds to androgen receptors and thus

serves a function of a growth hormone. Epitestosterone is an inactive epimer of testosterone.

In addition to analytical tests, a Bayesian statistical approach, similar to Sotas, has been mentioned and studied in many circles as a way to profile individual athletes and scrutinize specific samples that fall out of their normal percentile range. The objectives of this research are two-fold: one objective is to take two Bayesian approaches (the known-sigma model and Gibbs sampling algorithm) and build an athlete-specific model that can profile and alert an analyst if any potential doping occurs. The other objective is to build a model that is also inclusive of an athlete's specific genetic and biochemical makeup; that is, an athlete who dopes might still show a T/E ratio below that WADA-mandated threshold of 4 and an athlete that does not dope might show concentrations above 4. This model will alert an analyst if an athlete dopes based directly off his historical trends according to his population's statistical characteristics.

Preliminary Work

For the past five years, the Sports Medicine Research and Testing Laboratory (SMRTL) has tested half the NFL player population. The tests have been conducted based on World Anti-Doping Agency (WADA) mandates, which are known to be the strictest in the world. Currently, there are 35 Wada-accredited labs throughout the world, three of which are in North America. Any urinalysis received from an athlete follows a consistent

methodical testing procedure as prescribed by WADA in order to ensure consistency and fairness throughout the testing process. Over the years, SMRTL has collected an extraordinary amount of data for various sporting agencies, including the NFL. With this in mind, the preliminary steps include collecting all retrospective NFL data between January 1, 2006 to December 31, 2008 and then cleaning it up. The difficulty with this stage is that endogenous anabolic steroid levels can vary significantly from individual to individual and there is no set protocol that definitively eliminates outliers. The data cleanup will be discussed in the data clean-up procedure.

The other objective is to see if certain distributions for the entire NFL population matched that for other published populations; fortunately, this was also confirmed, which will be elaborated later. Finally, Sotas et al.[2] discusses the merits of using Bayesian statistics in order to highlight unusual biomarkers in a longitudinal steroid model. Unfortunately, the specifics of the method developed by his lab are not published due to proprietary reasons. The objective of this research is to develop a similar Bayesian model that does the same thing as Sotas without any controlled starting point other than retrospective data from the NFL population. The disadvantage with this study is that the population is not controlled in any way other than the fact that all members are male and belong to the NFL.

Bayesian known sigma model

One proposed fixed equation model for this study is the known-variance Bayesian model. This model utilizes the refined database of 3,108 different

NFL players tested from January 1, 2006 to December 31, 2008. These data provide relatively good information on the distributional characteristics for current parameters that reflect key ratios and compound concentrations. Hence, the population information on these ratios and compounds required by the selective adaptive model –specifically the mean (μ), and variance (σ) – are readily available from the NFL population. The adaptive known-variance model assumes that the test results for a population are defined by the data ($x_1,x_2,....,x_n$) and form the distribution $N(\mu,\sigma^2)$ where μ is the known-mean of the population and σ^2 is the known-variance of the population data. In addition, the player's distribution (prior distribution) is defined as $N(\nu_o, \tau_o^2)$ for some specified choices of ν_o and τ_o^2, both of which are unknown[6]. In other words, if there is nothing known about a specific player from the NFL population, then the statistics from a decile component of the NFL population will be used initially based off the player's first test result. As new data from the athlete become available, his statistics gradually dominate the model. From the given data, the player's expected posterior mean and variance are calculated from the following formulas:

$$E(\mu \mid y) = \tilde{\nu} = \frac{\tau_o^2 n\bar{x} + \sigma^2 \nu_o}{\sigma^2 + n\tau_o^2} \qquad \text{Posterior Mean}$$

$$V(\mu \mid y) = \tilde{\tau}^2 = \frac{\sigma^2 \tau_o^2}{\sigma^2 + n\tau_o^2} \qquad \text{Posterior Variance}$$

The posterior mean is a weighted average of the prior mean and the sample mean. The posterior variance is also based on a weighted calculation of the population and sample variance. To establish a reasonable Bayesian posterior percentile for a particular player's posterior mean and a weighted average of the overall population's standard deviation with the player's standard deviation, the 95^{th} and 97.5^{th} posterior percentiles are included based on standard protocols used in toxicology. The 95^{th} and 97.5^{th} percentiles are arbitrary depending upon the need to avoid any false identification of positive test results. In this instance, the model will simply be used to identify a player's results for further potential testing only and not for sanctions.

Gibbs Sampler Model

Another Bayesian approach is to use a Markov chain Monte Carlo algorithm and approximate a posterior statistical value from given joint conditional distributions. The mean, μ, is the variable of interest where $\mu=(\mu_1,\ldots,\mu_p)$. The joint posterior distribution of μ, denoted by $[\mu|\text{data}]$, is difficult to summarize and draw from since little information about the athlete is initially given. The set of conditional distributions is given as

$$[\mu_1|\mu_2,\ldots, \mu_p, \text{data}],$$

$$[\mu_2|\mu_1,\mu_3,\ldots, \mu_p, \text{data}],$$

……

$$[\mu_p|\mu_1,\ldots, \mu_{p-1}, \text{data}]$$

The idea is that an individual's corrected T/E ratios can be used to setup a Markov chain simulation algorithm from their joint posterior distribution by simulating from a set of p conditional distributions. Each individual parameter from these distributions represents one cycle of Gibbs sampling. In essence, samples generated from the Gibbs sampling algorithm will converge to their target distribution.[7]

One precursor for this course of action is to ensure that individuals in this population have lognormal distributions throughout both modes. Table 2 summarizes the results of various athletes within the distribution where the last three athletes (marked with *) tested positive for a synthetic anabolic steroid and player 969 tested positive for corrected T/E. Most of these distributions are parameterizable which affirms that the distribution for each athlete can use the natural conjugate prior, which, in turn, means that statistical parameters are represented by one distribution as oppose to different distributions for each parameter.[8] With this observation, a simplistic approach is to use the Gibbs sampling protocol to map a visible pattern that is indicative of steroid doping once it occurs. The data reflect a conjugate prior where the normal-inverse gamma distribution NIG(μ_o, c, a, b) reflects the player's unknown distribution based on patterns that exist with his peers. Once his data become available with each new test, then the normal-inverse gamma distribution yields to a posterior normal-inverse gamma distribution NIG($\tilde{\mu},\tilde{c},\tilde{a},\tilde{b}$) where

$$\tilde{\mu} = w\bar{y} + (1-w)\mu_o$$

$$\tilde{c} = w/n$$

$$w = nc/(1+nc)$$

$$\tilde{a} = a + n/2$$

$$\tilde{b} = b + SS/2$$

$$SS = (n-1)s^2 + (w/c)(\bar{y} - \mu_o)^2$$

where, $\bar{y}$ is the sample mean and s^2 is the sample variance. Naturally, this is too difficult to do in Excel, but WinBugs has this function built in. The code utilized in WinBugs is listed in the Appendix.

Table 2: Individual Player Distributions

		Data Characteristics				Distribution Parameters		Goodness of Fit Tests at α=0.05 Reject Ho?		
Player	RSD(CV)	min	max	n	Distribution	σ	μ	Kolmogrov-Smirnov	Anderson-Darling	Chi-Squared
975	250.74%	0.1	3.86	16	Lognormal	0.22551	-2.0014	No	No	No
4357	150.24%	0.11	1.73	21	Lognormal	0.54076	-1.7654	Yes	Yes	Yes
26545	125.67%	0.07	0.78	13	Lognormal	0.62669	-2.1872	No	No	No
23250	106.22%	0.53	5.41	20	Lognormal	0.47867	-0.19899	No	No	Yes
1011	91.40%	0.14	1.18	18	Lognormal	0.43153	-1.3509	No	No	No
25482	82.39%	0.14	0.96	15	Lognormal	0.4733	-1.5481	No	No	No
25481	77.15%	0.4	2.17	10	Lognormal	0.50308	-0.50269	No	No	N/A
4485	71.00%	0.93	4.53	10	Lognormal	0.43096	0.28906	No	No	No
2933	66.16%	1.21	7.29	18	Lognormal	0.38233	0.61276	No	No	No
4352	65.12%	1.16	5.87	11	Lognormal	0.43213	0.59669	No	No	No
20442	62.35%	0.89	4.27	11	Lognormal	0.41184	0.3126	No	No	No
719	62.09%	0.92	4.3	11	Lognormal	0.42304	0.32946	No	No	No
748	60.95%	0.63	3.4	16	Lognormal	0.38179	-0.03701	No	No	No
5345	60.40%	1.24	5.02	10	Lognormal	0.46024	0.59281	No	No	N/A
22402	59.13%	0.49	2.17	10	Lognormal	0.394	-0.29004	No	No	N/A
23241	59.09%	0.58	2.63	9	Lognormal	0.4038	-0.05914	No	No	N/A

3214	54.72%	1.38	6.05	10	Lognormal	0.41491	0.81667	No	No	N/A
3844	52.53%	0.5	1.99	10	Lognormal	0.39774	-0.24447	No	No	N/A
24891	52.35%	0.13	0.71	14	Lognormal	0.3922	-1.4107	No	No	N/A
1498	52.30%	0.1	0.45	10	Lognormal	0.40631	-1.743	No	No	N/A
21419	51.86%	1.15	5.01	17	Lognormal	0.39931	0.64698	No	No	No
24916	51.82%	0.63	2.62	14	Lognormal	0.34973	-0.10699	No	No	No
24899	48.68%	0.13	0.51	12	Lognormal	0.39172	-1.5878	No	No	No
687	48.58%	0.17	0.84	27	Lognormal	0.34365	-1.2897	No	No	No
1743	48.38%	0.87	3.84	13	Lognormal	0.33976	0.34688	No	No	No
1807	48.22%	0.1	0.41	23	Lognormal	0.36289	-1.8526	No	No	No
22805	48.04%	0.21	0.81	11	Lognormal	0.40842	-1.0193	No	No	No
4367	47.58%	0.53	2.03	12	Lognormal	0.60148	-5.06E-05	No	No	No
23345	47.44%	1	4.33	19	Lognormal	0.30473	0.35751	Yes	Yes	No
7641	46.88%	0.09	0.37	10	Lognormal	0.41023	-1.7701	No	No	N/A
5922	46.69%	0.06	0.23	8	Lognormal	0.3683	-2.271	No	No	N/A
25012	45.53%	1.12	4.81	13	Lognormal	0.41582	0.57186	No	No	No
986	45.04%	0.12	0.41	12	Lognormal	0.32967	-1.7886	No	No	No
1810	44.41%	0.81	2.82	11	Lognormal	0.38317	0.34109	No	No	N/A
21400	44.31%	0.52	2.05	13	Lognormal	0.32835	-0.19803	No	No	N/A
20461	44.28%	0.84	3.58	17	Lognormal	0.31184	0.26064	No	No	N/A
4416	42.77%	0.68	2.35	13	Lognormal	0.33589	-0.0192	No	No	No
2257	42.25%	0.5	1.82	14	Lognormal	0.3575	-0.16808	No	No	No
1831	42.07%	2.22	7.77	14	Lognormal	0.34656	1.2255	No	No	No
24863	41.78%	1.32	4.55	14	Lognormal	0.28216	0.58757	No	No	N/A
21715	40.95%	0.12	0.56	20	Lognormal	0.30868	-1.5691	No	No	No
23215	40.93%	0.6	2.4	18	Lognormal	0.34964	0.02603	No	No	No
24864	40.60%	0.65	2.3	17	Lognormal	0.33337	-0.03468	No	No	No
3823	40.47%	1.35	4.08	10	Lognormal	0.3386	0.70742	No	No	No
23117	30.74%	0.66	1.58	8	Lognormal	0.25997	-0.0205	No	No	N/A
21390	30.65%	1.03	2.9	13	Lognormal	0.25756	0.42832	No	No	No
23147	30.12%	0.64	1.9	23	Lognormal	0.2762	0.05664	No	No	No
21576	30.08%	1.35	3.18	12	Lognormal	0.24031	0.52381	No	No	No
2284	29.93%	1.22	3.51	13	Lognormal	0.25329	0.61738	No	No	No
21442	20.69%	0.09	0.19	13	Lognormal	0.18145	-2.1312	No	No	N/A
2880	20.67%	0.93	1.69	13	Lognormal	0.19571	0.2026	No	No	No
23623	20.63%	0.83	1.57	13	Lognormal	0.1931	0.14099	No	No	No
22661	20.60%	0.68	1.37	11	Lognormal	0.20999	0.05001	No	No	No
23175	20.60%	1.16	2.51	14	Lognormal	0.19276	0.51472	No	No	No
21424	20.59%	1.66	3.3	15	Lognormal	0.18276	0.73224	No	No	No
21733	10.34%	1.21	1.71	10	Lognormal	0.09448	0.32401	No	No	No
23151	10.13%	0.81	1.09	10	Lognormal	0.09471	-0.06747	No	No	N/A
21603	9.59%	0.62	0.84	12	Lognormal	0.09097	-0.33847	No	No	No
772	9.53%	0.89	1.18	12	Lognormal	0.10064	-0.01515	No	No	No
24842	9.43%	0.66	0.88	9	Lognormal	0.08818	-0.29457	No	No	N/A
23140	9.43%	1.05	1.41	9	Lognormal	0.08963	0.21202	No	No	N/A
3611	8.84%	0.61	0.8	12	Lognormal	0.08294	-0.37697	No	No	No

1777	8.45%	0.59	0.76	10	Lognormal	0.08235	-0.38021	No	No	N/A
2322	8.40%	1.2	1.59	11	Lognormal	0.0834	0.27535	No	No	N/A
26871	8.11%	1.76	2.31	10	Lognormal	0.07804	0.70404	No	No	N/A
24844	7.60%	3.14	3.77	8	Lognormal	0.12462	1.2235	No	No	No
2643*	153.87%	2.07	34.87	8	Lognormal	0.9488	1.3858	No	No	N/A
24875*	16.64%	0.67	1.17	11	Lognormal	0.15504	-0.11652	No	No	No
969*	73.92%	0.59	5.03	30	Lognormal	0.35859	-0.05297	Yes	Yes	Yes
23117*	32.82%	0.59	1.29	10	Lognormal	0.30187	-0.01258	No	No	No

The distributions for seventy players were analyzed. The players spread across a broad spectrum where the difference between their lowest and highest corrected T/E results was the factor taken into consideration for their distributional analysis. The last four players marked with (*) tested positive for some synthetic anabolic steroid. Most players have a lognormal distribution with two parameters.

METHODS AND MATERIALS

Chemical Extraction Procedure

The urinalysis and determination of synthetic anabolic steroid abuse follows a set protocol as mandated by the World Anti-Doping Agency (WADA). The initial sample is 3mL of urine taken randomly from an athlete at an unknown time. The urine sample is first tested for pH and specific gravity in order to determine if the athlete has attempted to dilute his urine prior to a drug test. Once the physical characteristics of the urine are tested, then the chemical procedure ensues. The extraction process is as follows: the sample is first buffered and hydrolyzed with β-glucuronidase, re-extracted with methyl tert-butyl ether, dried down and derivatized with n-methyl-n-trifluoroacetamide. The sample is then analyzed using an Agilent 7890A GC/MS-SIM. The peaks generated are compared to a known standard with

known concentrations based on WADA protocols; for any sample peaks that match retention times and meets or exceeds standard quantification, then that sample is marked as a potential positive. The confirmation process for synthetic endogenous steroids is to re-extract the urine sample, but then use ion-ratio mass spectrometry (IRMS) and examine the generated C-13 to C-12 ratio. Synthetic endogenous steroids tend to have fewer C-12 atoms than their endogenously produce counterparts, which means that a C-13/C-12 ratio greater than 3 is indicative of synthetic anabolic steroid doping according to WADA.[9]

Data Cleanup Procedure

The original dataset of NFL samples collected from January 1, 2006 to December 31, 2009 contains N_o = 17759 samples. The variables within each sample are as follows: the player identification number, sample date, pH, specific gravity, T/E, corrected T/E, testosterone, epitestostrone, DHEA, androsterone, etiocholanolone, and d3-testosterone/d3-epitestosterone. In addition, the ratios androsterone/etiocholanolone and androsterone/testosterone were manually calculated and added to the available variables. The difficulty with this data set is primarily that all data available in the database were manually entered by analysts employed at Sports Medicine Research and Testing Lab (or SMRTL) based on the results obtained from the GC-MS system. The other problem is that the results obtained from the GC-MS system could easily be incorrect if the analyst who

performed the evaluation neglected to calibrate the instrument or adjust the appropriate retention times. In essence, there are a lot of possibilities of data error in the data set and it is very difficult to retrieve archived data and rebuild the exact analytical GC-MS method to match the original parameters in which the dataset was originally generated. Variables such as temperature, column type, column length, compound consistency, and even the age of the instrument are all nominal and difficult to precisely duplicate, which prevent a perfect reenactment of the original analysis, so the data can not be validated after a certain point and unfortunately the data used for this paper fall out of time range where any meaningful correction can occur. In order to compensate for this issue, a cleaning methodology was agreed upon and applied consistently throughout the dataset. Even though the objective of this specific project is to model corrected T/E ratios, other endogenous anabolic steroids were included in the clean-up since these will also be used to build a multivariate Bayesian model during a later stage of funding. The cleaning methodology included entire sample deletions as well specific variate deletions within each sample with the overall goal to keep as much data as possible. First, 224 samples were completely removed since they had too many missing variables and would yield too little information to carry them to any further stage. Second, a specific gravity range was determined to be [1.00, 1.04], where 1.00 is the specific gravity of non-iodized water; any points above or below this range were removed –685 samples were removed following this guideline. Specific gravity, even though not directly

incorporated in the model, must be known in order to normalize all endogenous steroid concentrations for a given athlete. All variates used in future models, other than ratios, will be normalized in order to mediate the concentration variation for endogenous anabolic steroids based on nominal urine density. Third, corrected T/E and T/E given in the data set should both exceed 0. To ensure its integrity, corrected T/E is calculated directly from the T/E and the d3-testosterone/d3-epitestosterone ratios, also given in the database. D3-testosterone/d3-epitestosterone ratios, it must be noted, act strictly as an internal standard or control in order to ensure that the ratios from one sample to the next follow a law of proportions that can easily be violated by any abnormalities related to instrumentation or the extraction process. The following formula computes corrected T/E:

$$\text{Corrected T/E} = (\text{T/E}) \times 4 / (\text{d3-testosterone/d3-epitestestosterone})$$

Theoretically, the calculated values should match the given corrected T/E in the database. This test was carried out further by calculating the absolute difference between the given corrected T/E and directly calculated corrected T/E. Any values that exceed 0.25 were noted and removed from the dataset. Overall, 438 samples fall out of this range. Another data characteristic to note, values for testosterone less than 10ng/ml are not quantitated due to limits with instrumentation and analysis methods. This also applies to epitestosterone where the value is less the 2ng/ml. Corrected T/E ratios are

automatically calculated once the data is processed by the analyst; however, the difficulty remains validating this data since it is impossible to calculate a ratio with two unknown values. One final insurance step was to note and remove variable outliers using a bootstrap method. For the bootstrapping procedure, the upper fourth quartile of the data poll was examined, mainly the upper limit of the 99th percentile. For each compound, 1000 iterations were used in order to determine the upper limit. This coefficient was then multiplied by 2 and that value was used as the outlier cutoff point. Table 3 summarizes the output from the bootstrap method. Note that most samples fell within range, only a handful of samples were removed using the bootstrap method.

Table 3: Bootstrap Summary

Parameter	Mean	Std. Dev	Outlier Factor	Outliers Removed	New Mean	New Std. Dev
Corrected T/E	1.289	1.041	12.360	1	1.286	1.006
Testosterone	35.4	20.5	264.0	1	35.3	20.0
Epitestosterone	38.3	24.9	322.0	1	38.3	24.3
Androsterone	2725	1378	18730	0	2725	1378
Etiocholanalone	1930	997	14094	0	1930	997
DHEA	35.5	21.5	302.0	0	35.5	21.5
Andro/Etio	1.57	0.87	9.96	8	1.56	0.72
Andro/Test	100.4	65.9	668.3	1	100.3	65.4

Mean and standard deviation refer to the statistical parameters before any bootstrap determined outliers were removed from the data set. The new mean and standard deviation indicate the new statistical parameters once outliers were removed.

Fourth, all endogenous anabolic steroids were corrected based on specific gravity and dilution factors. One strategy that athletes may use to thwart a

urinalysis is to dilute their urine. Specific gravity is measured for each sample before the actual urinalysis begins and any sample that has a specific gravity less than 1.01 is doubled. This translates into dividing each endogenous steroid concentration by 2 if they had a corresponding specific gravity less than 1.01 since those values are based off twice the volume of urine. Fifth, specific gravity was normalized and that value was used to normalize testosterone, epitestosterone, etiocholanolone, DHEA, and androsterone:

$$\text{Normalized Specific Gravity} = (1.02-1)/(\text{raw specific gravity} - 1)$$

Finally, all positive samples are noted within the dataset. As for this dataset there are four positive corrected T/E ratios for three different athletes; in addition, there are positive samples that entirely consists of one boldenone metabolite, three methyltestosterone metabolites, one epimetenediol, and one epitestosterone at 601. With all the above correction factors there are N=15736 samples that contain all information needed for the overall analysis. However, with the exception of the 224 samples listed in part 1, no other samples have been removed in their entirety. There are 16,934 negative samples where corrected T/E is greater than zero and four additional corrected T/E samples where the player tested positive. Table 4 shows the corrected T/E summary for all negative samples and this will be the variable used to build the preliminary models.

Table 4: Corrected T/E Statistics for All Negative Athletes

Variable	N	Mean	Std. Dev	Variance	Min	50th %	95th %	Max
Corrected T/E	16934	1.30	1.01	1.03	0.02	1.07	1.30	11.69

Summary statistics for corrected T/E for all negatively tested athletes. This summary does not include any positive tests. Note: there are 416 corrected T/E values greater than 4.0, which, according to WADA, are indications of synthetic anabolic steroid doping; however, all these values were confirmed negative by IRMS.

CHAPTER 3

RESULTS

Data characteristics

There are multiple variables examined and the distribution characteristics differ widely between some compounds and yet are similar between others: the distributional summaries have been included for completeness. The focus of this paper is on corrected T/E ratios, but eventually, the other variables will be included in a multivariate model since they all represent endogenous anabolic steroid levels. It must be noted that the corrected T/E distribution was the only variable that formed an observable bimodal distribution, which is expected with this dataset.

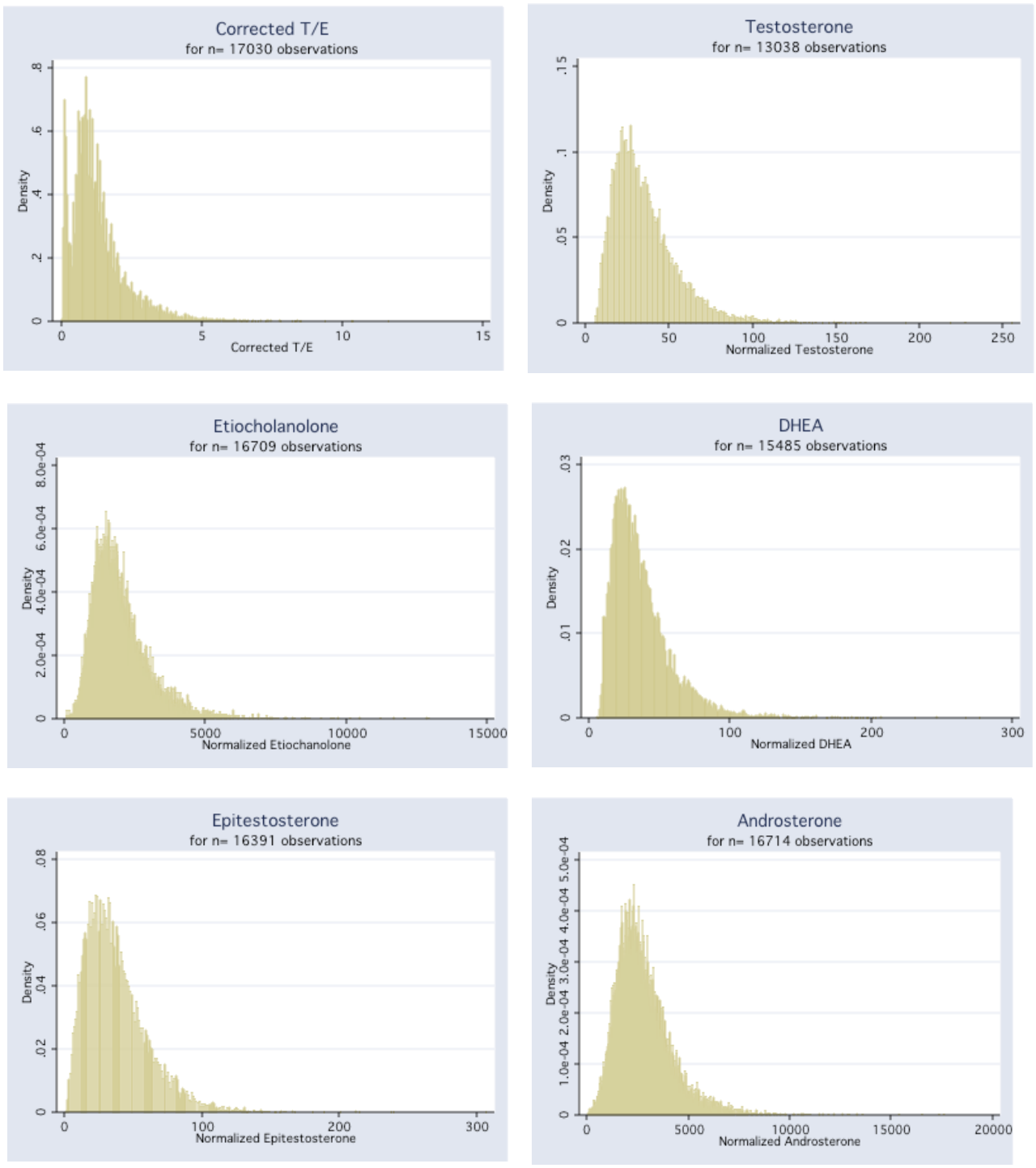


Figure 2: Histogram for non-positive NFL players from January 1, 2006 to December 31, 2008. Corrected T/E is composed of both testosterone data and epitestosterone data. The bimodal distribution as evident in the histogram for corrected T/E should be seen with the testosterone, but LOQ restraints did not record the first mode for testosterone.

Note the bimodal distribution of T/E, which is indicative of a minority population that carries UGT 2B17 double deletion polymorphism.[10] This bimodal characteristic and its relative proportions of "low" and "normal" basal corrected T/E ratio has been well documented by other groups.[3 11 12 13] This bimodal distribution is not seen in the graph of testosterone concentrations since the limit of quantification (LOQ) of the assay used for these analysis is 10 ng/mL, which excluded values below that threshold. The other distributions are included since there are no known results of distributions that can be compared to them. Figure 3 shows a box plot for all athletes that have more than 20 tests within the given time reference.

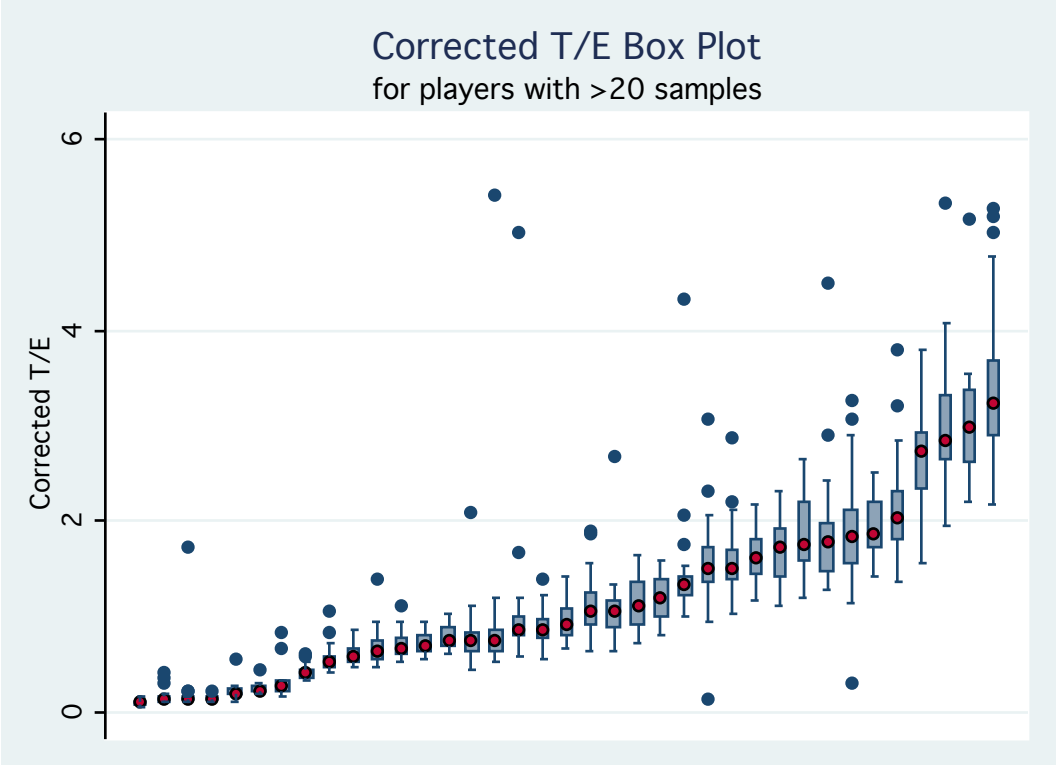


Figure 3: Box-plot for players that had more than 20 tests from January 1, 2006 to December 31, 2008. All positive tests have been removed from the data set.

The ranges vary significantly with most athletes, but this is to be expected given the erratic behavior in which testosterone levels vary based on the previous explanations. It must be noted, however, that all these points represent negative tests, even those above the 4.0 threshold.

A Bayesian Model for a Positive Athlete

Two players were chosen that had sufficient data points to give a broad picture of the overall model: one tested positive for testosterone doping and one did not. Table 5 shows the general statistical characteristics for these two players.

Table 5: Statistical Summary for Two Individual Players

Player	Mean	St Dev	N	Min	10^{th} Percentile	Median	95^{th} Percentile	Max
Positive (with positive test included)	1.049	0.775	30	0.59	0.75	0.865	1.68	5.03
Positive (without positive test included)	0.912	0.193	29	0.59	0.74	0.86	1.20	1.68
Negative	0.88	0.176	30	0.55	0.67	0.87	1.22	1.39

Table 5: Statistical summary of two players (one that tested negative for steroid doping and the other that tested positive for testosterone doping) that have an equivalent number of data points.

Figures 4 and 5 show the results for the known-sigma model and the Gibbs sampling algorithm, respectively, for a player that had a confirmed positive corrected T/E test at n=10.

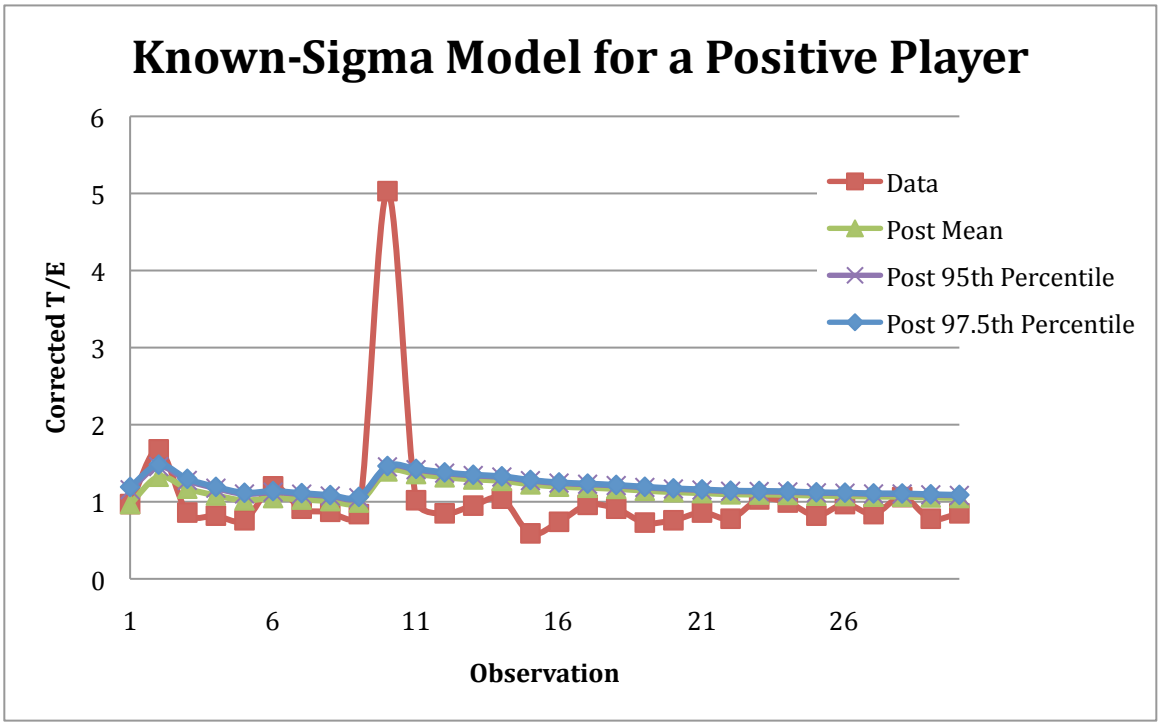
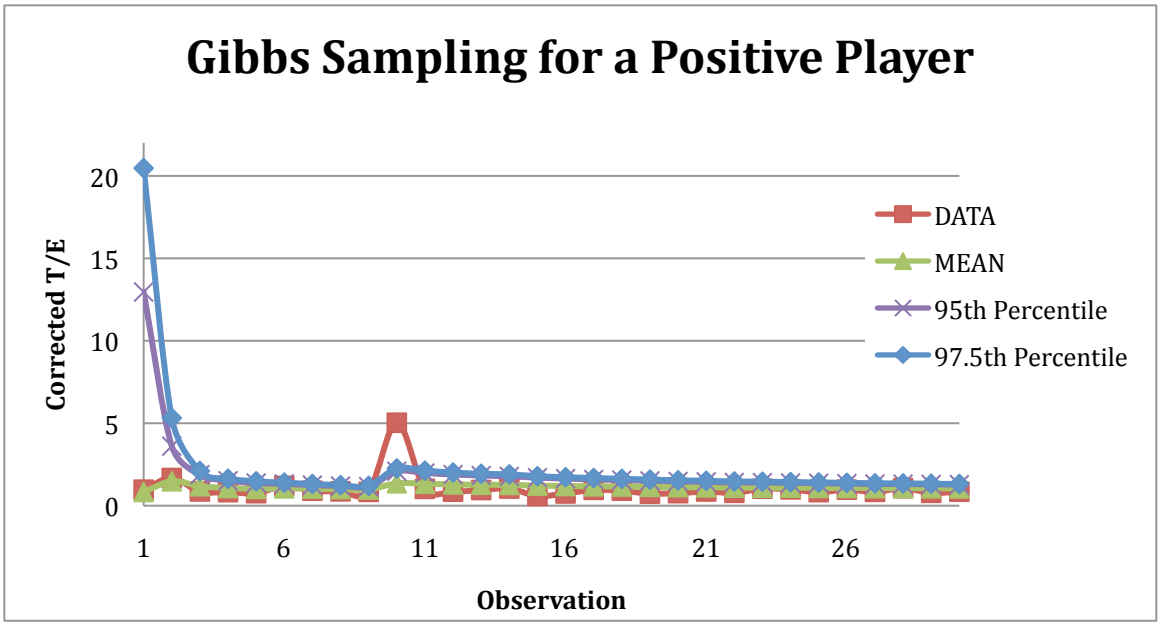


Figure 4: Data for a NFL player that tested positive at n=10. The known-sigma model is the basis for this output, which shows both the posterior mean and the posterior percentiles.



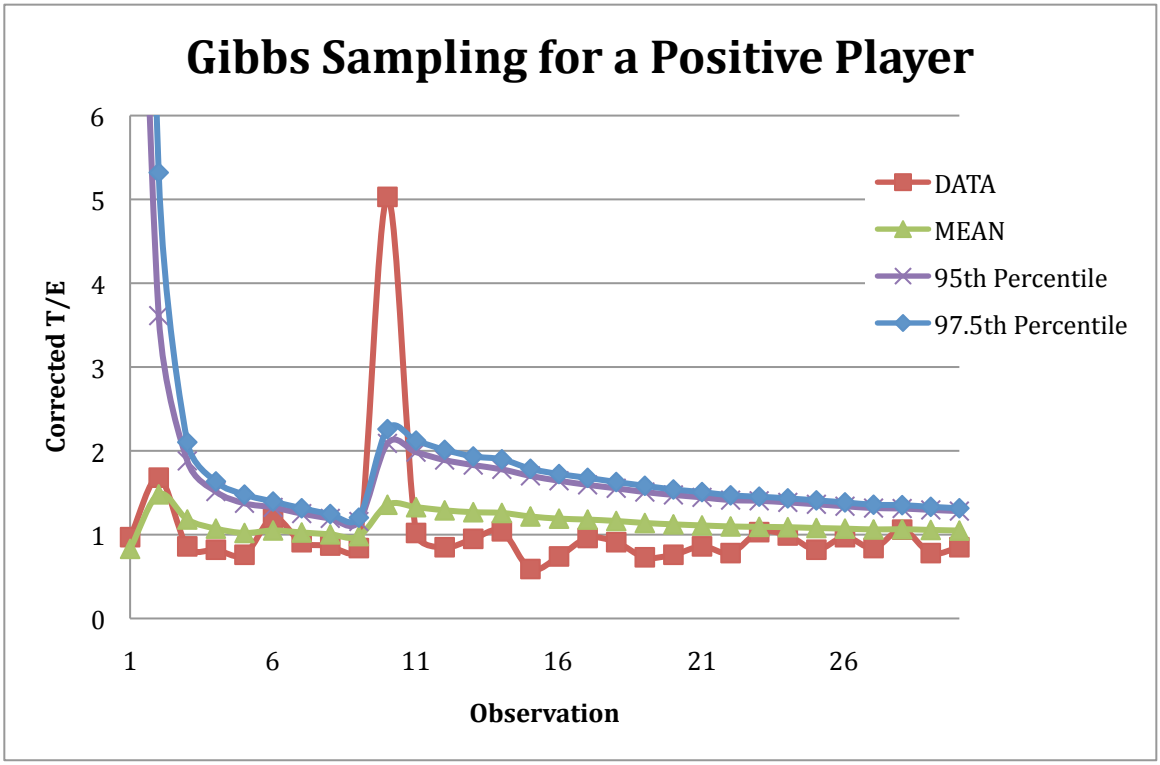


Figure 5: The first graph shows the Gibbs sampling algorithm and where the initial points begin. The second graph is drawn on a similar scale to that of the known-sigma model in order for fair comparison. Here the data and mean are below both the 95th and 97.5th posterior percentiles with the exception of point n=10, where this player tested positive for testosterone doping.

The known sigma model from figure 4 has a tighter initial range mainly since the population variance for that athlete in his respective decile is relatively tight (population variance is 0.0128). The prior mean for the player was set at one, which is relatively conservative, but his prior variance is set relatively high at 1000. The advantage of the known-sigma model is that if population information is known, which is the case for this data set, then the player's prior information has very little weight in the overall model, which is evident by the tight initial range for his first corrected T/E data point. The one confirmed positive point clearly violates both the 95th and 97.5th posterior percentiles and would clearly indicate to an analyst that there is some irregularity with this specific data point. In addition, both the 95th and 97.5th posterior percentiles do not dramatically increase with the positive test, but stay relatively close to their pre-positive levels.

Figure 5 shows the Gibbs sampling approach, which, in many ways mirrors the known-sigma model. There is a clear violation at n=10, which is evident in the model by the size of its jump and also it violates both the 95th and 97.5th posterior percentiles. One noticeable feature of this model is that the initial three points have high posterior percentiles, which suggest that this model has a high level of uncertainty with the initial points. The same population variance and player mean were used in the initialization of the model, so the same preconditions are utilized in both the known-sigma and Gibbs sampling models, but the Gibbs sampling procedure has a steeper learning curve than the known-sigma model. Another observation is that

both 95th and 97.5th posterior percentiles have significant lag adjusting back to their pre-positive levels as evident in figure 5.

A Bayesian Model for a Negative Athlete

Figures 6 and 7 show both the known-sigma model and the Gibbs sampling algorithm for a negative player. Again for the known-sigma model, the player's respective population decile variance is 0.0128 and his prior mean is one and his prior variance is 1000. As observed with the positive-tested player, both the 95th and 97.5th posterior percentiles rapidly adjust with the initial point and the majority of points fall below both intervals. There are slight variations in which the data are recorded above the posterior percentiles (for points n=3, 5, 12, 17, 19, 21, 22, and 27) and this does show an imperfection with this model (the same problem exists with the Gibbs sampling algorithm which we shall see shortly). An explanation is that testosterone levels for an individual can vary significantly based on many biological factors and, unfortunately, the testing procedures are never static. An approach for SMRTL is to develop the testing method so that there is less variation due to instrumentation issues. However, the data range for this model is not significantly outside the posterior percentiles and none of these points would have raised any red flags since they are all below the WADA cutoff of 4 for corrected T/E.

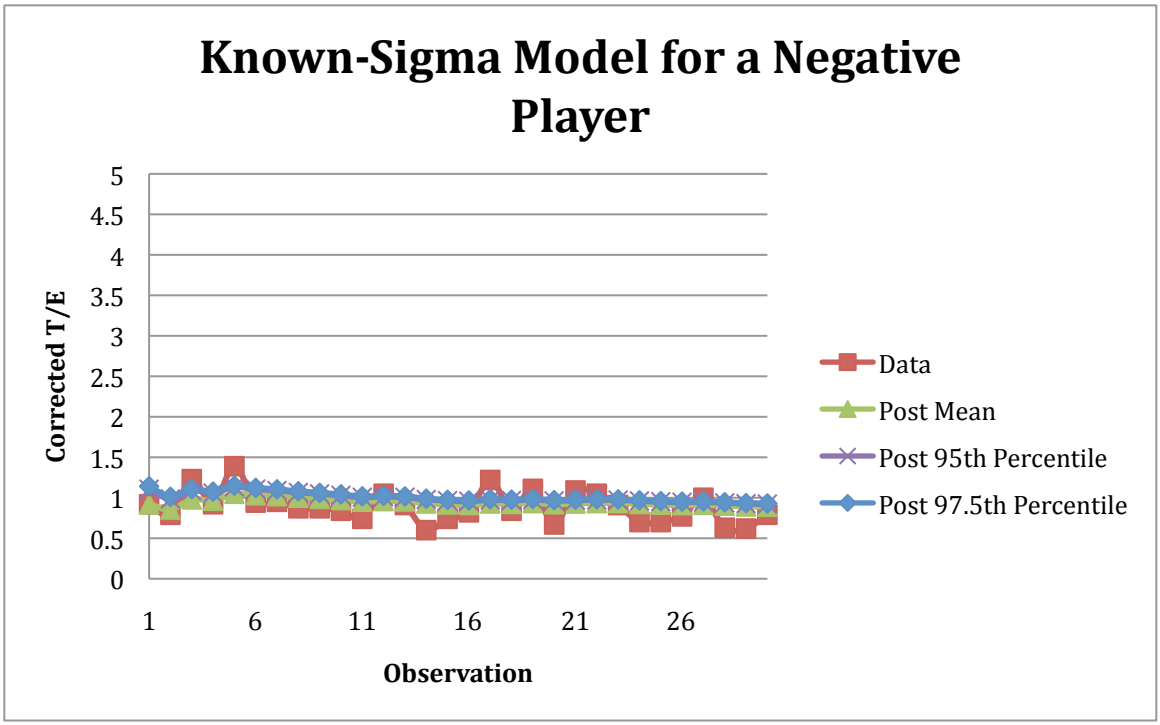
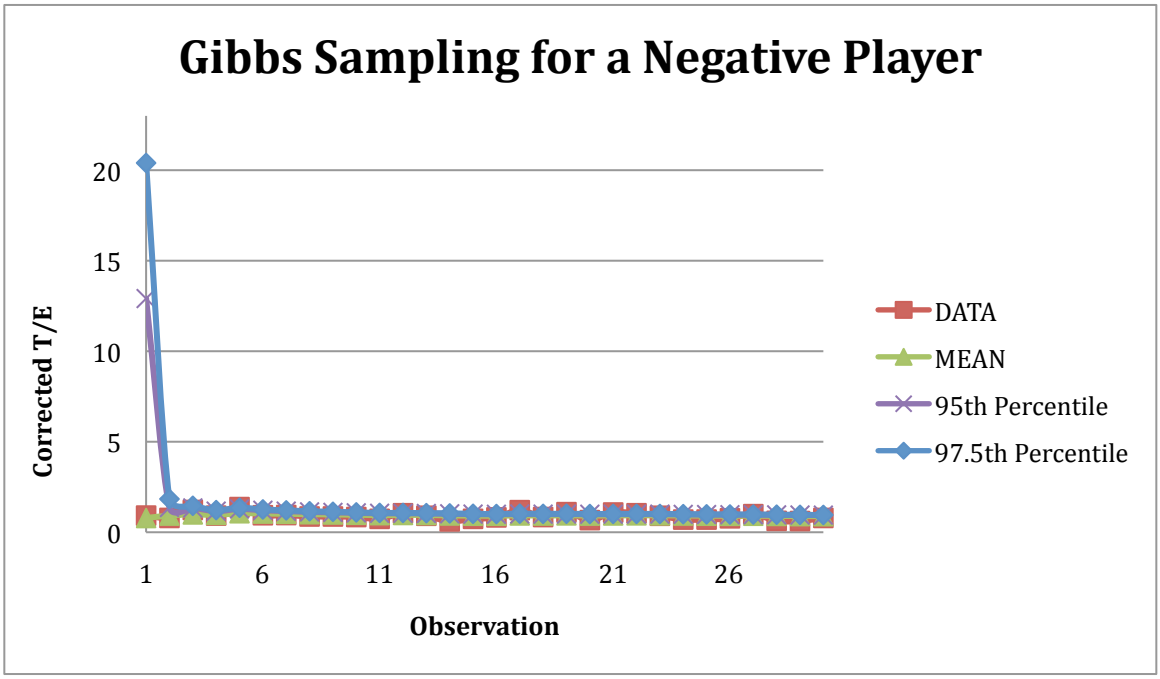


Figure 6: The known-sigma model for an NFL player that was never tested positive for corrected T/E from January 1, 2006 to December 31, 2008.



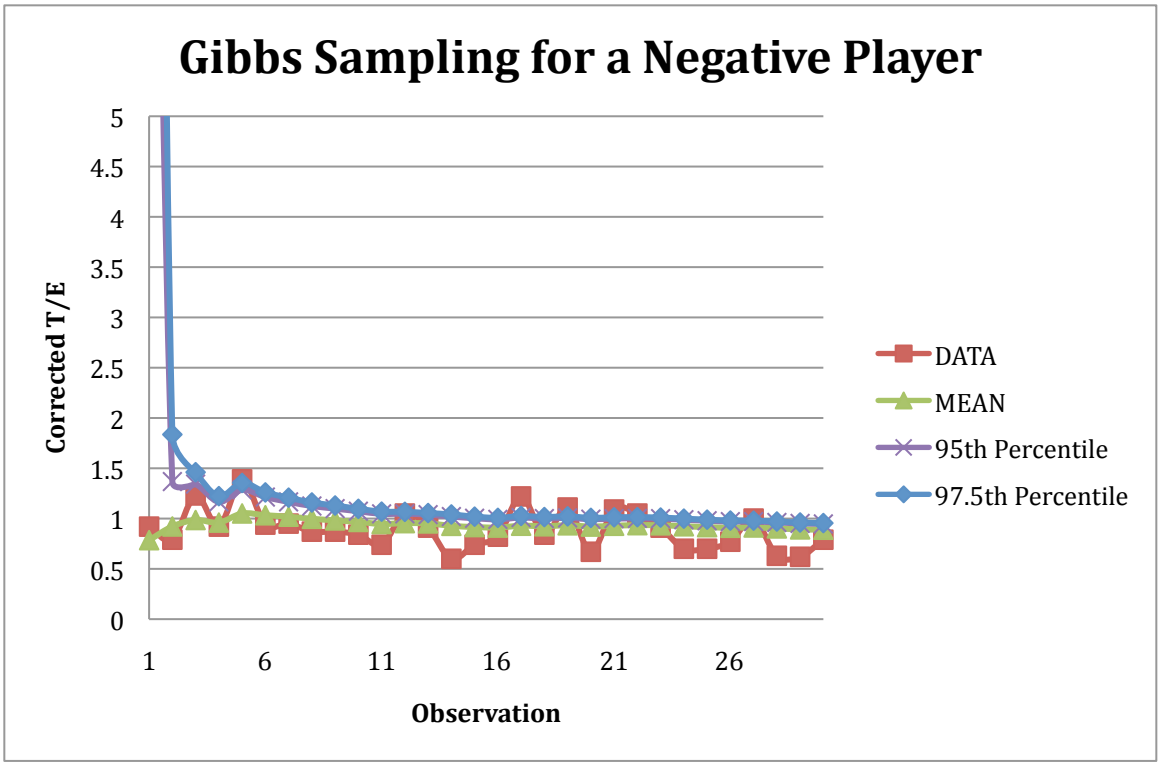


Figure 7: The first graph shows the Gibbs sampling algorithm where the initial points begin. The second graph is drawn on a similar scale to that of the known-sigma model in order for fair comparison. Here the data and mean are below both the 95th and 97.5th posterior percentiles with the exception of points n=5, 12, 17, 18, 20, 22, and 26.

Figure 7 shows the Gibbs sampling algorithm for a negative player. Similar to figure 5 for the positive-tested athlete, there is still a huge variation for the initial three points and the model behavior is very similar to the known-sigma model for the negative tested athlete in figure 6 with one noticeable exception: the known sigma model shows eight points that exceed the 97.5th posterior percentile whereas the Gibbs sampling algorithm shows only seven points above the 97.5th posterior percentile (points n=5, 12, 17, 18, 20, 20, 22, and 26). This may seem trivial with this data set, but it does seem to suggest that the Gibbs sampling algorithm manages to contain the majority of

negative data below their posterior percentiles, although it can contain the first three points due to the steep initial learning curve of its posterior percentiles. It must be noted that the Gibbs sampling procedure mirrors the model output from Sotas et al[3].

Quantification

The data cleanup and the determination of statistical attributes of the overall NFL population were conducted using STATA 11.0. The equations for the known-sigma model were used in Excel. WinBugs was utilized to determine all statistical parameters for the Gibbs sampling algorithm.

CHAPTER 4

DISCUSSION

Significance of findings

The first key finding is that the bimodal distribution of corrected T/E matches that of previous measured distributions.[3] Another key finding is that given a general NFL population with known statistical parameters, it is indeed possible to build a player-specific model that shows the appropriate posterior percentile based primarily on the population's mean and variance with the initial point, but as more player information becomes known, then the player's posterior mean and variance begin to dominate the model. The premise of this model can be applied to every athletic demographic where

members of that sport have physical and possibly genetic characteristics similar to other members within that specific athletic discipline, but might not have the same characteristics compared to members of another athletic discipline. As long as there is enough information specific to a population, primarily variance, then it is possible to build such models across all athletic disciplines with known population statistics.

The Gibbs sampling also provides an adequate model for modeling corrected T/E ratios for individual athletes. The similarity of this approach is that population information is as critical for this model to function as to that of the known-sigma model. The only additional property that has to be determined is if members in the given population had parameterizable distributions that would allow us to justify using a conjugate approach building the Gibbs algorithm. Similarly, for this study the population was divided into deciles where the player's initial corrected T/E was used to determine in which potential decile he belonged. The variance and mean of the population at the determined decile were used as the initial point for the Gibbs sampling algorithm. This approach is debatable since it does not seem to matter much where the initial point is drawn. Experimentation with other initial points seem to draw the same conclusion after two or three of the player's data points were entered into the model. Another notable feature of the Gibbs sampling method is that it matches the output provided by the Sotas et al paper.[3]

Known-Sigma model versus Gibbs sampling algorithm model

Both models have adequately modeled two separate athletes where one tested positive for testosterone doping at one point and another never tested positive for testosterone doping within the same given time interval. The known-sigma model does use the population's variance in which a given athlete is a subset. The initial point is heavily dependent on the population parameters, but as more information about the athlete is known over multiple testing rounds, then his posterior distribution becomes increasingly dependent upon his own information as seen from formulas 1 and 2. Although the athlete in this case is a member of the NFL and the population variance used consists of only NFL players, there is still significant non-specificity with a given athlete. As mentioned earlier, testosterone concentration is dependent on race, age, weight, overall health, and genetic characteristic as evident from the bimodal histogram in figure 2. In reality, there are still many unknowns with respect to this population and where the athlete fits in. Dividing the distribution into deciles is one method to remedy this problem, but there still could be overlapping subsets that form a specific decile in which the athlete may only belong to one subset. However, this problem will only exist with the initial tests since the athlete's statistical parameters begin to dominate the model as more of his tests become available.

The Gibbs sampling algorithm, a member of a Monte-Carlo Markov Chain procedure, too, is dependent on an athlete's population statistical

characteristics. Even though the model was initialized using the decile procedure mentioned above, experimenting with similar values for mean and variance did not change the overall behavior of the model much as far as the initial point is concerned. The initial three testing points have significant uncertainty related to posterior percentiles, but as more information becomes available the athlete's statistical parameters begin to dominate the model. The pattern almost mirrors that of the known-sigma model with the exception that if a positive test occurs, then the Gibbs sampling model seems to require more player data in order to adjust the posterior percentiles to their pre-positive levels.

In reality, both models do the same thing over long periods of testing; however, the Gibbs sampling algorithm shows the level of uncertainty with the first three testing points since so little information of the player is available. The known-sigma model adjusts immediately with the first three testing points and remains relatively consistent throughout the testing cycle. By observing the fact that one player from a population distribution, a distribution which is composed of many possible biological variables, is unknown suggests that the first three points of the known-sigma model should reflect the uncertainty of the population and, therefore, look similar to that of the Gibbs sampler approach. This suggests that the Gibbs sampling algorithm is a more realistic model for steroid behavior for a given athlete within this population.

Limitations and future work

One possible way that an athlete could trip up this model is to dope with his first and second tests. If the analyst misses these two positive tests, which can easily be done, then the threshold for the model will be too high, especially if the athlete continues doping throughout his athletic career. Another limitation is that this model does not include any information related to true positives. Calibration of this model may require an athlete to willingly dope at certain time points within his career; this would assuredly demonstrate the model's sensitivity and specificity with regard to the tested athlete. A future funding request will involve testing for this fact by having athletes within a certain profile willingly dope with testosterone.

Future work with respect to this model is to take all variables within the dataset and build a multivariate adaptive model. Since biological metabolism and pathways are seldom insulated, it would be practical to conclude that that doping for one compound could easily raise multiple red flags with respect to the other compounds. Additional compounds such as 5-α-androstenediol and 5-β-androstenediol will also be added to the model since the ratios of these two compounds have direct correlation with endogenous anabolic testosterone levels.

Conclusions and Relevance to Toxicology and Economics and other fields

The known-sigma adaptive Bayesian model and the Gibbs sampling approach developed based on NFL data can be applied across all athletic programs. As

long as each program has enough data to generate known statistical parameters, the known-sigma model and the Gibbs sampling algorithm can be implemented. Characteristics of these models' methodology are their robustness across disciplines including other areas of toxicology. For instance, it could be used to alert health professionals of individuals exposed to toxins that live within a radius of an industrial center. The adaptive Bayesian approach with known sigma and the Gibbs sampling algorithm can easily be applied to medicine, some examples include the following: the development of a model for healthy people according to BMI variations; PSA concentrations in older men and incidence of prostate cancer; or even estrogen levels and incidence of stroke. Some examples in economics include the monitoring of red flags at a bank that might trigger an audit; a specific market irregularity that might be a prelude to a crash; or even a social pattern that might indicate the results of an election. Both models can work with all of the above problems with the limitation that both models have to have a known population in order to initialize properly.

CHAPTER 5

APPENDIX

WinBugs Code

Gibbs Sampling Model

```
Model{
For(i in 1:n)
{
x[i]~dnorm(mu, tau);
}
xnew~dnorm(mu, tau);
mu~dnorm(1, 0.001);
tau~dgamma(0.001, 0.001);
sigma<-1/sqrt(tau);
xbar<-mean(x[]);
}
```

Initial Point

List(mu=1, tau=6103.516)

CHAPTER 6

REFERENCES

[1] Ehrlich, Isaac. Crime, Punishment, and the Market for Offenses. *Journal of Economic Perspectives* Winter 1996; 10: 43-67.
[2] Yesalis, Charles. *Anabolic Steroids in Sport and Exercise*, (2nd ED.). New York, NY: Human Kinetics. Pages 15-41
[3] Sottas P, Baume N, Saudan C, Schweizer C, Kamber M, Saugy M, Bayesian detection of abnormal values in longitudinal biomarkers with an application to T/E ratio. *Biostatistics* 2007; 8: 285-296.
[4] Dewick, Paul. *Medicinal Natural Products: A Biosynthetic Approach,* (2nd ED.). New York, NY: Wiley. Pages 282-283.
[5] WADA, in WADA (Editor). Guideline, The World Anti-Doping Agency, 2006.
[6] Ntzoufras, Ioannis. *Bayesian Modeling Using WinBugs*, (1st ED.). Hoboken, NJ: Wiley. Page 10.
[7] Albert, Jim. *Bayesian Computation with R* (2nd ED.). New York, NY: Springer. Page 122.
[8] Ntzoufras, Ioannis. *Bayesian Modeling Using WinBugs* (1st ED.). Hoboken, NJ: Wiley. Pages 42, 71-74.
[9] Shackleton CHL, Roitman E, Phillips A, Chang, T. Androstanediol and 5-androstenediol profiling for detecting exogenously administered dihydrotestosterone, epitestosterone, and dehydroepiandrosterone: Potential use in gas chromatography isotope ratio mass spectrometry. *Steroids* 1997; 62:665-673.
[10] Jakobsson J, Ekstrom L, Inotsume N, Garle M, Lorentzon M, Ohlsson C, Roh H-K, Carlstrom K, Rane A, Large Differences in Testosterone Excretion in Korean and Swedish Men Are Strongly Associated with a UDP-Glucuronosyl Transferase 2B17 Polymorphism. *J Clin Endocrinal Metab* 2006; 91: 687-693
[11] Van Renterghem P, Van Eenoo P, Geyer H, Schanzer W, Delbeke FT, Reference ranges for urinary concentrations and ratios of endogenous steroids, which can be used as markers for steroid misuse, in a Caucasian population of athletes. *Steroids* 2010; 75: 154-163.
[12] Catlin DH, Hatton CK, Starcevic S. Issues in detecting abuse of zenobiotic anabolic steroids and testosterone by analysis of athletes' urine. *Clin. Chem*. 1997; 43: 1280-1288.
[13] Ayotte C, Goudreault D, Charlebois A. Testing for natural and synthetic anabolic agents in human urine. *Journal of Chromatography B: Biomedical Sciences and Applications* 1996; 687: 3-25.