Evaluating Metrics for Comparing Powerlifters

Brant Imhoff[1] and Michael J. O'Connell[1]

[1]Department of Statistics, Miami University, Oxford, OH 45056, U.S.A.

1 Abstract

The sport of powerlifting has sought for a way to best represent the relative strength of its lifters. Various scores have been developed with the goal of representing its lifters fairly with respect to weight class, biological sex, and equipment status. Robert Wilks' score derived from a fifth-order polynomial regression using a small sample, the Wilks Score, has been the most agreed upon score across powerlifting for the last three decades. Powerlifters have recently suggested amending the Wilks Score or adopting another score that more fairly represents lifters across aforementioned demographics. Other authors have proposed solutions that involve models simpler than Wilks' fifth-order polynomial to prevent overfitting. This analysis generates multiple polynomial regression models trained on the current IPF sample (N=59,047) to develop novel scores. To compare the newly created scores from this analysis and the Wilks Score, differences in χ^2 goodness-of-fit statistics are used to evaluate how well these scores fit the IPF data relative to each other. A nonparametric bootstrap technique is used to approximate the variability in goodness-of-fit differences between powerlifting scores. Quantile-based confidence intervals are created to measure effect size of goodness-of-fit differences. It is shown that the P2 score, based on a second-order polynomial regression, provides a noticeably better fit for bootstrapped IPF samples than the Wilks Score, IPF Points, and other novel scores developed in this analysis. Alternative methods of analysis and the necessity of updating powerlifting scores is also explored.

2 Introduction

Competitive powerlifting is a strength sport in which an individual attempts to lift the maximum amount of weight possible. A competition is comprised of three different types of lifts in the following order: squat, bench press, and deadlift [7, 2]. For each of these three lifts, an individual is given three separate attempts to lift as much as possible in that respective lift [7, 2]. A powerlifting competition gives each lifter nine total attempts and the goal is to have the highest total amount of weight lifted, taking the sum of the heaviest performed squat, bench press, and deadlift.

Competitors are divided into weight classes, which are separated by sex [7, 3]. This separation into weight classes coincides with the idea that the more mass an individual holds, the more mass that person should be able to lift. Additionally, classic powerlifting, which is also known as raw powerlifting, competes separately from equipped powerlifting. Raw powerlifting limits a lifter to gear that is not significantly assistive in nature along with the uniform all lifters must wear (an approved singlet, shoes, socks, and a t-shirt). Per IPF (International Powerlifting Federation) raw powerlifting rules, a raw lifter is able to use a supportive belt, knee sleeves, and wrist wraps [7, 10-15]. Equipped lifters are allowed to wear gear that provides more compression and benefits the lifter more than typical lifting gear. These items include a squat/deadlift suit under the singlet, knee wraps (which are different from knee sleeves used in raw lifting), and a bench press shirt [7, 10-12]. Raw and equipped lifters do not compete against one another in conventional powerlifting competitions, even if they fall into the same weight class. Thus, there are not only weight class divisions, but equipped and raw divisions. For example, males that fall into the 93 kg weight class either compete in the 93 kg raw division or 93 kg equipped division. In addition to weight classes, competitions also allow lifters to enter age divisions. There are teen divisions, junior and collegiate, the open division which is the largest and where anyone between 24 and 39 is required to compete, and masters (40+) divisions [7, 3].

To determine the best lifter within each weight class, it may seem simple at first: the

person with the highest total kg at the end of the competition wins the weight class. This would be the solution if everyone were to weigh the same amount. It has been suggested by research that strength is not necessarily a linear function of bodyweight. Astrand and Rodahl suggested that strength and body mass have a relationship that would be better characterized by a polynomial function [3]. This presents the idea of scaling total kg lifted in powerlifting competitions by a competitor's body mass. It also of interest to do this as fairly and representative as possible. Robert Wilks, the inventor of the Wilks Formula, provided the best solution to this problem with a set of metrics derived from fifth-order polynomial regression equations [20]. His motivation for creating the Wilks Formula was to best characterize the relationship between body mass and strength in powerlifting [5]. The Wilks Formula gave lifters the ability to compare themselves across weight classes. It also allowed for comparisons between sexes, providing better recognition for female lifters in a sport where males are the majority. This formula has been used by all federations as its standard for measuring strength for decades but has been critiqued in recent years for its biases against middleweights and favorable biases for heavyweights and lightweights. Because of this the IPF moved to a new metric called IPF Points, which assumes a Log-Normal distribution of bodyweight [15]. IPF Points gives each lifter a score based on his/her log body weight and equipped/classic status, then adds a constant such that there are larger gaps between scores to make comparisons easier [6]. Other metrics such as Wilks and the ones developed in this analysis also scale based on similar variables, just with different functional forms.

2.1 The Wilks Formula and Overfitting

Roberts Wilks last updated his fifth-order polynomial equations in 1995, suggesting no difference in coefficients from his initial model that was fit in 1984 [5]. The primary concern with the Wilks Score is its model form. Polynomial regression equations with high order coefficients, typically anything above a quadratic or cubic term, start to take on extreme

values [9]. It is likely that the Wilks equations were overfitting the concurrent powerlifting data. Models that overfit are overestimating the complexity of the relationship between the predictors and the response. Graphically, this would be represented by a fitted curve that is not smooth enough for the data which it is attempting to approximate. Higher-order polynomials that overfit typically perform exceptionally well when evaluating goodness-of-fit to the data on which the models were trained, but perform noticeably worse on new data [14]. It would be unfair to say that the Wilks Score has not been useful. Rather, its use has only come under scrutiny as of late because of the growth of powerlifting. More people are beginning to question if the Wilks Score is the right fit for the powerlifting community. Its use for selecting the best lifter at an elite competition may make sense due to the extreme values a fifth-order polynomial would produce, mimicking the behavior of world-class powerlifters. However, the complexity of a fifth-order polynomial is highly questionable for modeling the relationship between mass and strength for anyone who is not internationally recognized, which is the vast majority of lifters. Thus, this analysis explores other statistical models with smoother functions to create novel powerlifting scores. Additionally, a fifth-order polynomial equation using updated, higher-dimension training data will be used to mimic the Wilks Formula. The goal of building these models is to best explain the relationship between a lifter's total and their demographic characteristics and body weight and compare novel scores to the Wilks Score, as well as determine how and where the Wilks Formula overfits.

2.2 Biases of Modern Powerlifting Scores

Bias in the most general sense refers to an unfair or favorable inclination. Statistical bias uses the same framework to describe a point estimator of a population parameter. A point estimator is any function of a sample. The idea of using a point estimator is to try to best estimate a population parameter in the interest of gaining knowledge about the population, finding useful interpretation of parameters, or approximating a function of a parameter.

There are a few desirable qualities of point estimators, often depending on whether the sample is small or sufficiently large enough to deploy asymptotic properties. For use in this paper, the estimator criterion of interest is unbiasedness. A statistically unbiased point estimator is able to, on average, precisely estimate the parameter of interest. The formal definition of bias is as follows, providing an explicit mathematical definition of unbiasedness:

Equation for Bias of a Point Estimator

$$Bias_{\theta}(\hat{\theta}) = E(\hat{\theta}) - \theta$$

$\hat{\theta}$ is a point estimator for the parameter θ

An estimator is said to be unbiased if its bias is 0, suggesting its expectation is equal to the parameter of interest. Smaller bias indicates that an estimator is close in approximating the true parameter which it is attempting to estimate [4]. In most scenarios, unbiasedness is a desirable property of estimators, as it suggests that the estimator is accurate. An applied discussion of unbiasedness for sample proportions in powerlifting immediately follows.

Minimizing bias of the powerlifting scores is imperative. People pay money to participate in powerlifting, whether that be in the form of membership fees, meet fees, equipment, or travel expenses. Federations should honor the contributions and dedication of their members by using the fairest scoring method possible, even if in the grand scheme of it all, most people are not competing for titles or prize money. Fair representation by weight class, sex, age, equipment choice, etc. will continue to grow the sport. Perhaps the most discussed reason for maintaining fairness in scores is the selection of the best lifter at competitions. When titles and prize money are at stake, it's important to those who are competing that the right person is chosen. These reasons highlight the importance of fairness for the whole community and top lifters.

Prior to discussing methods to evaluate potential biases of the original Wilks formula, an updated Wilks formula, and the scores developed in this study, it is imperative to define

Table 1: Classification of IPF, USAPL Lifters by Percentiles [18]

Lifter Classification	Elite	Master	Class 1	Class 2	Class 3	Class 4	Class 5	Novice
Quantile	.975	.95	.85	.75	.5	.25	.1	0

unbiasedness in this context. An unbiased powerlifting formula would rank lifters such that representation by demographic would be the same in smaller samples as demographic representation in the full powerlifting community. More specifically, the proportion of lifters within a certain classification/skill level (seen in Table 1) for the entire IPF given a certain demographic or variable (ie. sex, weight class, equipment) should be approximately the same as the proportion of lifters in that same demographic for any given sample from the IPF. This idea of unbiasedness should apply particularly to the covariates that were adjusted for in statistical models, which are body weight (which should translate to weight class), sex, and equipment. A statistical representation of this concept of bias is illustrated in the equation below.

Equation for Bias of an Estimator of IPF Proportion

$$Bias_{Metric}(\hat{p}) = E(\frac{d}{C}) - \frac{D}{IPF} \tag{1}$$

$$= E(\hat{p}) - p \tag{2}$$

d is the number of lifters from a specified demographic such as weight class, sex, or equipment choice for a specified classification, limited to only lifters in **C**. **C** is the sample size of a specific classification/skill level (Table 1) of lifters based on ranking from a metric. **D** is the number of lifters from a specified demographic such as weight class, sex, or equipment choice within the entire IPF sample. **IPF** is the number of total lifters in the IPF sample, which is 59,407 in this analysis.

Given this definition of bias, if the expectation of the proportion of a specific demographic of lifters within a sample matches the proportion of that same demographic in the entire IPF, the method of estimation is statistically unbiased for that specific sample. An example for this analysis would be that for a specific metric, Wilks perhaps, all lifters are ranked

by Wilks and the top 2.5% (Elite lifters) are taken, then the proportion of Elite lifters who are in the 105 kg. weight class is the estimated proportion. If the proportion of Elite 105 kg. lifters matches the true proportion (for the whole IPF) of 105 kg. lifters, then Wilks would be unbiased in that specific instance for the 105 kg. weight class. For this analysis, bias will be evaluated for different models and compared using a test we describe later. This evaluation takes place after each model is developed into a score which is applied to each lifter where the lifters are then ranked by each metric, then divided into samples by lifter classification quantiles (see Table 1) based on their lifting abilities according to each metric.

Previous discussions of bias have been focused on how the distribution of de facto 'elite' lifters matches the entire lifter population, whether that be the top 100, top 1%, etc. Per IPF and USAPL standards, an elite lifter would be anyone in the top 2.5% of the federation when ranked by score (Wilks, IPF Points, P5, etc.). This idea of comparing elite lifters to the rest of the IPF can also be applied to the other six lifter classifications based on quantile values (see Table 1). If the distribution of the top lifters should match the IPF distribution, so should the distribution for novice lifters, as well as everyone in between. IPF Points was able to do this relatively well in Marksteiner's Proposal, as all weight classes by sex and equipment follow approximately Normal distributions [6]. The issue with IPF Points, though, comes in its model assumptions regarding the distribution of body weight. Marksteiner suggests in his proposal that in modeling a lifter's total or bench press, a lifter's body weight requires a natural logarithm transformation [15]. His model also provides different coefficients for sex and equipment. Justification for using the natural log of body weight, Marksteiner asserts, is that body weight is Log-Normally distributed by sex [6]. This assumption appears to be met graphically via Log-Normal quantile plots in Marksteiner's IPF Points Proposal [15, 14]. The same assumption was evaluated for the updated IPF data in this analysis, but was nowhere near being met despite following Marksteiner's filtering criteria and updating with more recent meet results (see supplementary materials).

In Validation of the Wilks powerlifting formula, bias was evaluated graphically by scat-

terplots of body weight with separate curves for each sex of the Wilks-adjusted totals and individual lifts [20]. If the points were distributed symmetrically about the fitted lines, then it was determined that Wilks was an unbiased measure of strength for that lift/total by sex for the body weight ranges represented in the data. This worked well because of the small subsample used in Validation of the Wilks powerlifting formula, which included 27 female lifters and 30 males. Thus, it was trivial to determine if there was bias a scatterplot given that there were so few data points around the fitted regression lines. It was determined that Wilks was unbiased for total kg. lifted given that the data were roughly uniformly distributed about the fitted regression line [20]. More recent works have suggested the Wilks Formula should be updated due to bias [12] and changes in weight classes and overall athlete body composition [8]. In this paper, evaluating biases is not as simple given the amount of data (about 60,000 unique IPF-affiliated lifters). Additionally, the samples of interest are the various lifter classifications (Elite, Master, Class 1, etc.) based on quantiles of total kilograms lifted or metrics such as IPF Points (see Table 1).

3 Methods

3.1 Data

The IPF sample was obtained from the open source database Open Powerlifting [17]. This database is maintained and updated regularly by a group of developers with the help of powerlifting enthusiasts. Data for this project were downloaded in .csv format from Open Powerlifting. The initial data set included approximately 1.4 million observations, each representing a competition for a powerlifter. Competition data were compiled from September 1964 to April 2019. Lifters had each of their attempts recorded, as well as their best attempt in each of the three lifts to form a total. Negative values for attempts indicated missing a lift, while positive values indicated a successful attempt. Additionally, measures such as body weight in kg., biological sex, equipment, and age were collected. For each row, scores

such as Wilks and IPF Points were given.

The scope of the initial data set was narrowed to include only IPF lifters given the interest of comparing Wilks, IPF Points (which was created using only IPF lifter data), and statistical models developed in this study that were trained only on IPF data. These data were filtered to include lifters in only IPF-affiliated federations and current IPF weight classes for full powerlifting meets, excluding push/pull and bench only competitions. Competition results with missing values for body weight, biological sex, and equipment were excluded. The categories of single-ply and multi-ply from the Equipment variable were reduced to one equipped category in a new variable EQ. The new EQ variable included only two categories: classic/raw and equipped. This was done in the interest of simplifying statistical models without losing much information given the IPF does not currently allow multi-ply equipment [7, 10]. The final IPF sample contained 59407 observations and 38 variables.

Variables of interest for statistical modeling include those for which bias will be evaluated: weight class, sex, and equipment. Strength can be evaluated on the basis of each of the aforementioned variables. In Figure 1 a density plot of total by weight class displays approximately bell-shaped, unimodal distributions across all weight classes. Although, variability appears to be higher in the male weight classes.

Weight classes are mutually exclusive on the basis of sex, so males and females do not compete against one another. Equipped and raw/classic lifters do not compete against one another, but the Wilks Score does not reflect the differences between these lifters. IPF Points operates on the assumption that equipped and raw/classic lifters have different distributions of strength, so their coefficient estimates are different. Figure 2 displays density plots for total demonstrating a difference in distribution between classic and equipped lifters for the IPF sample of nearly 60000 lifters.

Both distributions are bimodal, but the equipped lifter sample looks to have more variability than the classic/raw sample. Equipment should benefit a lifter by providing additional strength, which is what differentiates equipped lifting from raw lifting. This idea in addition

Figure 1: Density of Total by Weight Class

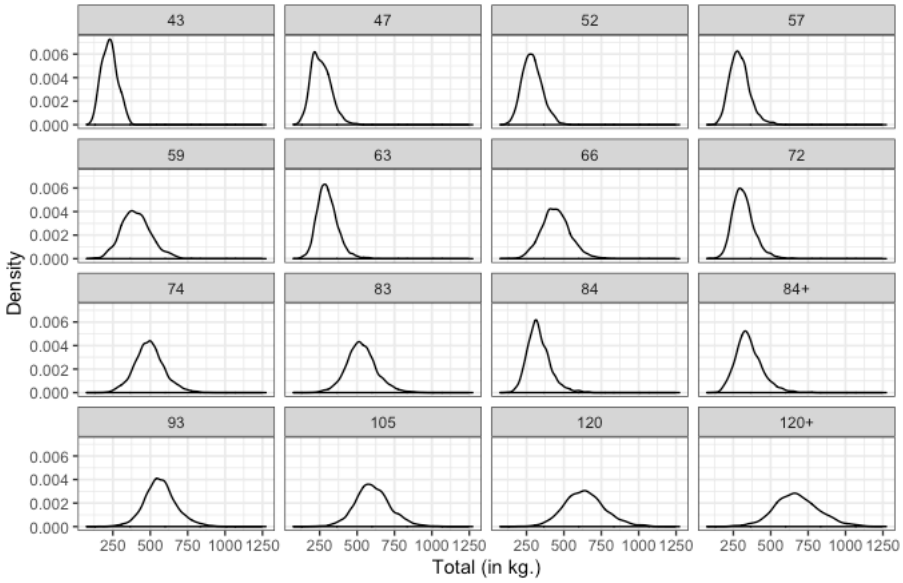
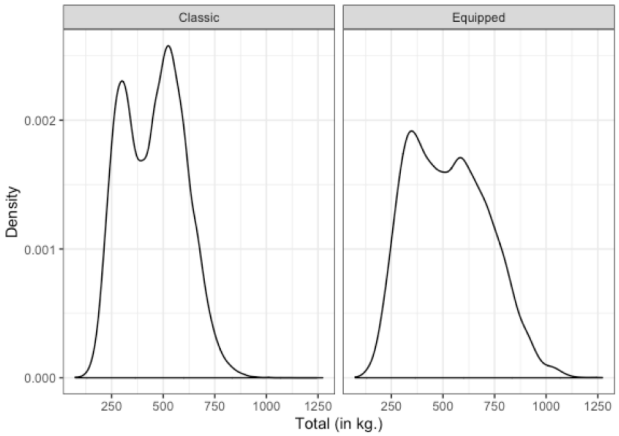


Figure 2: Density of Total by Equipment



to the difference in the distribution of strength (in the form of total kg. lifted) between equipped and classic/raw lifting provides justification for exploring statistical models that differentiate between equipped and classic lifters.

3.2 Models

The process of obtaining a powerlifting score that fairly represents lifters by varying demographic levels is a problem explanatory in nature. It requires a model form with extractable coefficients such that scores can be calculated simply via an equation or using a table of values. This is to best account for accessibility and availability of the powerlifting community so that anyone can calculate their individual score. Unfortunately, this constraint eliminates many of today's sophisticated statistical models and machine learning algorithms from contention in this analysis. Another requirement is that the model used to develop the score should adjust for the covariates that lifters desire. Particularly, these are body weight/weight class, sex, and equipment use. Other variables such as age are also considered, but are not as evident in their relevance to modeling a lifter's strength in the form of total kg. lifted. In candidate statistical models, it is also desirable if polynomial functions can be implemented. As previously mentioned, the relationship between body weight and strength is believed to be best approximated by a polynomial function [3]. Polynomial regression is an answer to all of the requirements mentioned above: extractable coefficients, inclusion of multiple covariates, and ability to include polynomial functions for body weight.

Polynomial regression meets the general requirements for modeling strength in powerlifting, is simple to deploy, and has favorable statistical properties. The first property is the ability to model non-linear relationships between predictors and the response [19], which is justified by literature in modeling strength with a polynomial function of body weight. However, the true functional form of the relationship between body weight and strength and is still unknown. This has led to different polynomial model forms to produce powerlifting scores, such as Wilks' fifth-order polynomial, Marksteiner's log body weight model,

and Khudayarov's fourth-order polynomial. These approximations of the true relationship between body weight and strength highlight another useful application of polynomial regression. When functions are unknown or difficult to approximate, polynomials may be used. The interest is not the polynomials, but how closely they approximate the true function and minimize error [19].

Candidate models, particularly those using fifth-order polynomials and log body weight terms, are automatically fit to determine if training the same models on more data influences goodness-of-fit when converted to powerlifting scores. In developing novel candidate models, certain criteria should be used for measuring how well the polynomial regressions approximate the true relationship and minimize error. Given the explanatory nature of this problem, the selection of novel candidate models relies on Adjusted R^2, testing Root Mean Squared Error (RMSE), and AIC. Testing RMSE was obtained via 10-fold cross validation using the caret package's train() function in R [13], while AIC and Adjusted R^2 were obtained from R's lm() function. Further, p-values of candidate model coefficients are used for guidance in deciding which polynomial terms are necessary. Procedures regarding polynomial approximations often involve fitting higher-order polynomials and lowering the order if simpler polynomials fit just as well [19]. A natural starting point for candidate models seems to be fifth-order polynomials. An interest of this analysis is to see if updating the Wilks Score and IPF Points provide high performance by goodness-of-fit measures, so the first models for consideration were the Updated Wilks, Updated IPF Points, and P5 models (which are described below). Working backwards from the fifth-order polynomial P5 model, models with quadratic, cubic, and quartic terms for body weight are fit with the same demographic interaction terms (sex and equipment). Prespecified model forms (Updated IPF Points and P5) use two-factor interactions in their model equations to match complexity of existing model forms. Quadratic, Cubic, and Quartic polynomial models use three-term interactions between body weight, sex, and equipment. No improvement in Adjusted R^2 is observed from the quadratic model all the way up to the fifth-order model, remaining at

0.68. This suggests that the quadratic model explains the same amount of variation in total kg. lifted as the other more complex models. The leveling of Adjusted R^2 with increasing polynomial order is particularly informative, especially because Adjusted R^2 is a function of the number of predictors used in a model. Additionally, the difference in testing RMSE from the quadratic model to fifth-order model is negligible. The second-order model's RMSE of 90.96 is quite comparable to the fifth-order polynomial model's RMSE of 90.84, suggesting virtually no difference predictive accuracy. The only noticeable difference between polynomial models comes from AIC, in which more complex models are generally favored. The fifth-order model's AIC of 704985.1 was noticeably lower than the quadratic model's AIC of 705264.7. However, the performance of the quadratic polynomial model against higher-order polynomials by Adjusted R^2 and testing RMSE justifies its consideration for developing a powerlifting score.

The Wilks formula was originally developed with data that are no longer representative of modern powerlifters [6]. To update this formula, a fifth-order polynomial regression model was fit using newer data with the same regression equation as the Wilks formula, given below. The coefficients from this model were used to generate updated Wilks scores in the same fashion in which Wilks scales using a ratio of observed over predicted and multiplies by a constant. Additionally, another fifth-order polynomial regression model that adjusts for equipment was fit to compare to the Wilks and Updated Wilks metrics. This regression model included additive terms for body weight, age, sex, and equipment, as well as interactions. We will call this new score the P5.

The general form of the Wilks and updated Wilks regression models are displayed below, with total kg. lifted as the response variable. These models use polynomials for body weight in kg. up to the fifth-order and biological sex, an indicator variable, as well as interactions to model the mean behavior of total kg. lifted.

Wilks Polynomial Regression Equation

$$\hat{Total} = \beta_0 + \beta_1 S + \sum_{i=1}^{5} \beta_{i+1} W^i + \sum_{i=1}^{5} \beta_{i+6} W^i S$$

Total is in kg., W represents body weight in kg., S is an indicator for Sex where 0 = Female, 1 = Male

While updating the Wilks Formula is one potential remedial measure for biases, another is to utilize other covariates. Of the candidate models, the best model determined by Adjusted R^2, testing RMSE, and AIC was also a fifth-order polynomial regression, but with additional predictors. Wilks utilizes bodyweight as a predictor, and has two separate formulas: one for men, one for women. The alternative equation for what will be dubbed the P5 uses the same idea as Wilks but adjusts for equipment and sex. The polynomial regression equation for P5 is as follows:

P5 Polynomial Regression Equation

$$\hat{Total} = \beta_0 + \beta_1 E + \beta_2 S + \beta_3 ES + \sum_{i=1}^{5} \beta_{i+3} W^i + \sum_{i=1}^{5} \beta_{i+8} W^i E + \sum_{i=1}^{5} \beta_{i+13} W^i S$$

Total is in kg., W represents body weight in kg., S is an indicator for Sex where Male = 1 and Female = 0,

E is an indicator where Equipped = 1 and Classic/Raw = 0

In addition to the P5 model, two other models were developed to adjust for equipment, which is excluded from the Wilks formula. The P2 is derived from a polynomial regression model that includes a quadratic term for bodyweight in addition to covariates for sex and equipment. The Updated IPF Points metric utilizes log bodyweight instead of a linear term, but also adjusts for sex and equipment, following the same functional form as the model used to develop IPF Points. These models are described below:

P2 Polynomial Regression Equation

$$\hat{Total} = \beta_0 + \beta_1 E + \beta_2 S + \beta_3 ES + \sum_{i=1}^{2} \beta_{i+3} W^i + \sum_{i=1}^{2} \beta_{i+5} W^i E + \sum_{i=1}^{2} \beta_{i+7} W^i S + \sum_{i=1}^{2} \beta_{i+9} W^i SE$$

Total is in kg., W represents body weight in kg., S is an indicator for Sex where Male = 1 and Female = 0,

E is an indicator where Equipped = 1 and Classic/Raw = 0

Updated IPF Points Polynomial Regression Equation

$$\hat{Total} = \beta_0 + \beta_1 E + \beta_2 S + \beta_3 ES + \beta_4 ln(W) + \beta_5 ln(W)E + \beta_6 ln(W)S$$

Total is in kg., W represents body weight in kg., S is an indicator for Sex where Male = 1 and Female = 0,

E is an indicator where Equipped = 1 and Classic/Raw = 0

Once all models were fit, their fitted values for total kg. lifted were used to create new powerlifting scores. The statistical models created in this analysis predict a lifter's total based on the relationship specified in the model form, implying that scaling still must be completed. Scaling with respect to all demographic groups would be the most ideal way to turn model predictions into powerlifting scores and evaluate bias on the basis of those variables, but it cannot be used for two reasons. The chief reason for not scaling this way is that it would likely result in a favorable bias when evaluating the goodness-of-fit of novel scores. Second, other powerlifting scores such as Wilks and IPF Points were not scaled this way and would make comparisons unfair. Because of this, an ad-hoc method of scaling will be used. This problem of creating an adjusted score from a linear model is sparse in the statistical literature but is familiar to the powerlifting community. Methods used previously by Robert Wilks in his fifth-order polynomial regression conversion to the Wilks Score can be utilized here. The general form of his scaling method is often glossed over, with no explanation or justification. From a statistics perspective, the scaling appears strange. However, when coupling the context of the problem with statistical ideas, the idea becomes simple and intuitive. If the polynomial model is believed to explain the relationship between strength and some predictors of choice, then comparing actual results to the model's predictions of strength provides a relative measure of strength. In simpler terms, the ratio of a lifter's observed total to predicted total can be thought of as how many times stronger a lifter is than what the model suggests they should be. This ratio of relative strength makes practical sense if the model is believed to be truly explanatory of strength and also provides a solution to the issue of unfair scaling for different metrics. Thus, the general form of a powerlifting score is as follows:

$$Score \; = \; C * \frac{Total\ kg.}{\hat{Total}\ kg.}$$

C is a positive constant used to inflate the **Total kg.**:$\hat{\textbf{Total}}$ **kg**. ratio, providing further space between individuals' scores.

Total kg. is the actual total kilograms lifted for a given powerlifter. $\hat{\textbf{Total}}$**kg**. is the predicted total kilograms lifted by a statistical model for a given powerlifter

3.3 Using Goodness-of-Fit to Assess Bias

In this analysis, χ^2 goodness-of-fit statistics are used to measure bias of powerlifting scores. The χ^2 goodness-of-fit statistic is used to measure how closely the distribution of groups within a sample match a hypothesized distribution. The flexibility of the χ^2 statistic allows for the analysis of nominal and ordinal categorical variables, as well as discretized continuous variables in the form of counts by group. It also provides nonparametric alternatives to when certain parametric assumptions are not met.

The primary goal of this analysis is to determine if powerlifting metrics rank lifters in a way that matches the overall distribution of lifters, with respect to a demographic variable. This naturally provides a setup of χ^2 tables such that the row variable is the demographic variable of choice, while the column variable is the skill level assigned to a lifter based on a specific powerlifting metric. The interest is to make goodness-of-fit comparisons between the different powerlifting scores. Further, this can be viewed as a goodness-of-fit evaluation in that the aim is to determine how well the rows of the table match a null distribution, conditioned on the column variable. The ability to use the test statistic as a goodness-of-fit measure comes from specifying an expected count in each cell, which can be based on a hypothesized or known distribution. The observed count in the cell is based on empirical data, while the expected count is based on a known distribution. Thus, because the χ^2 statistic (displayed below) is a function of the squared difference between the observed and expected counts, it serves as a distribution matching measurement.

χ^2 Goodness-of-Fit Statistic Calculation

$$\chi^2 = \sum_{r=1}^{i} \sum_{c=1}^{j} \frac{(O_{ij} - E_{ij})^2}{E_{ij}}$$

O_{ij} denotes the observed count for the cell in the ith row, jth column. E_{ij} denotes the expected count for the cell in the ith row, jth column.

While bias and goodness-of-fit are inherently two different statistical concepts, the two are naturally intertwined in this analysis of powerlifting scores. The goal of this analysis is to quantify bias of powerlifting scores by evaluating differences in goodness-of-fit statistics. The parameter of interest to be estimated is the count of lifters within each cross-section of skill level (ranking quantile based on powerlifting score) and demographic. The count parameter is based on the population proportion of lifters who fall within that demographic of interest for the entire lifting federation. The point estimate of count is the observed number of lifters who fall into the skill level and demographic cross section given the powerlifting score used. Thus, if a powerlifting score is able to generate a point estimate for count that matches the count parameter based on the whole federation, the score produces an unbiased estimate. Goodness-of-fit statistics become useful in quantifying bias in this application for two reasons. First, there are many cross-sections (r x c) of lifter classifications (skill level) and demographic level for each table. A bias calculation becomes convoluted in this scenario and does not provide one summative measure for each powerlifting metric. Second, the goodness-of-fit statistic is a function of bias for each individual cell, given that it takes the difference between point estimate for count and count parameter (see χ^2 Goodness-of-Fit Statistic Calculation for a Powerlifting Score in Section 3.4). The χ^2 statistic takes this function of bias and sums it for the whole table, providing a summative measure of bias via the goodness-of-fit statistic for a specific powerlifting score. Comparing these statistics then allows for comparisons in bias between powerlifting scores.

Further justification in using the χ^2 test statistic for goodness-of-fit can be gathered from statistical literature. Authors such as Goodman [10] and Agresti [1] use χ^2 statistics as

goodness-of-fit measures for rxc contingency tables but rely on parametric models to provide alternatives to the χ^2 test itself. Similarly, these authors use a χ^2 goodness-of-fit statistic without performing a χ^2 test in their analyses. Unlike the aforementioned authors' analyses, the aim of this analysis is not to perform a hypothesis test that relies on a specified model form of goodness-of-fit statistics. Instead, the variability of the goodness-of-fit statistic is of interest. This leads to the implementation of the nonparametric bootstrap, which will be discussed in detail in the following section.

The ultimate goal of this analysis is to measure the goodness-of-fit to the full IPF sample of the Wilks Score, IPF Points, and novel scores generated from the models described in the previous section. However, the full IPF sample is dynamic, continually growing along with the strength of its competitors. If the IPF data weren't changing so frequently and requiring the updating of lifting scores to meet expectations of fairness, then the IPF sample of 59,047 lifters that trained the models developed in this analysis could be considered the true population. This would make the goodness-of-fit evaluation much simpler in practice, given each metric's performance is limited to this one sample.

Evaluating goodness-of-fit can be completed in a variety of ways. Jeong et. al. used the bootstrap method (discussed in Section 3.4) to conduct non-parametric hypothesis tests with χ^2 tables for two categorical variables [11]. Their computational process is similar to the one conducted here, commonly using statistical models fit to an initial sample, the implementation of bootstrap samples, and calculation of χ^2 test statistics for each bootstrap sample to evaluate goodness of fit of each metric. The difference between the methods used here and those used by Jeong et. al. is the way inference is conducted and that their χ^2 statistics are model-adjusted, based on the RC model proposed by Goodman [11] [10]. The analysis of powerlifting metrics relies on the comparison of the distributions of χ^2 test statistics and confidence intervals from bootstrap samples, while Jeong et. al. compressed their results to p-values [11].

3.4 Comparing Methods with a Bootstrap Test

Now that bias has been defined, the methods of evaluating biases for each metric may be outlined. The goal of developing an alternative to Wilks and IPF Points is two-fold. First, it is desired to provide a metric that more fairly distributes lifters regardless of classification amongst demographics, whether it be by weight class, sex, or classic/equipped status. Second, a metric should accomplish the aforementioned duty for different data sources that are representative of the IPF population, not just the IPF sample on which the models were trained. The abundance of data from the past few decades of powerlifting inherently has led to the development of models which are more flexible than Wilks' fifth-order polynomial. The popularity of the sport will also provide a stream of new data in the future. In an attempt to better approximate how fairly powerlifting scores rank IPF lifters, the statistical tool of the bootstrap may be used to quantify the variability of the goodness-of-fit of these metrics. The bootstrap involves taking the data at one's disposal and resampling it randomly with replacement, matching the same sample size as the original data. A sample statistic is then taken from each bootstrap sample. The goal of bootstrapping is to build a distribution around the sample statistic in order to better encapsulate the quantity's variability.

In this analysis, the sample statistic of interest is the difference in χ^2 goodness-of-fit statistics. Thus, this analysis will utilize bootstrap distributions of pairwise goodness-of-fit statistic differences between powerlifting scores. To summarize the information contained in these bootstrap distributions, quantile-based confidence intervals for differences in χ^2 goodness-of-fit statistics are included to demonstrate which scores provide the best fit to IPF data.

This procedure involving the bootstrapping of test-statistics may be classified as non-parametric. The χ^2 statistic is inherently nonparametric [16]. Though it is assumed that expected counts within χ^2 table cells are known, these are conditional upon each iteration of entire bootstrap IPF sample. When evaluating differences in χ^2 statistics, finding a solution for the distributional form becomes challenging. This eliminates the option of conventional

hypothesis testing and confidence intervals that rely on a distribution. The nonparametric bootstrap solves this problem by simulating this unknown distribution by sampling from the original sample with replacement and generating goodness-of-fit statistics for each iteration.

In practice, the idea is to take the 59,047 unique best totals of IPF lifters and bootstrap a large number of samples. For each bootstrap sample, each lifter is then ranked by each of the six metrics. Once ranked, lifters are divided into classifications (see Table 1) and for each weight class, the count of lifters within that classification/quantile is taken. This is completed for each of the six metrics, then a χ^2 table is generated for each of the six metrics with counts. The cross-section at each cell in the χ^2 tables refers to a sub-sample within each weight class and lifter classification for each of the six metrics. An example of this table may be found in the supplementary materials. Additionally, the steps utilized in this analysis to evaluate bias, quantify goodness-of-fit, and develop meaningful insights regarding powerlifting metrics are outlined below.

Steps of Computational Study

1. Develop Candidate Models
    (a) Select Candidate Models based on Adjusted R^2, 10-Fold CV RMSE and AIC
    (b) Generate Scores from Selected Models for all lifters
2. Perform Bootstrap Sample Computations
    (a) Generate Bootstrap 10,000 Samples from the Full IPF Data
    (b) For Each Bootstrap Sample, Assemble a χ^2 Table for each of the 6 Metrics
    (c) For each table, Calculate a χ^2 Test Statistic
    (d) Take the difference in χ^2 Test Statistics between Wilks and all other metrics
3. Inference
    (a) Generate a distribution of differences in χ^2 Test Statistics for comparison

(b) Create confidence intervals for differences in χ^2 Test Statistics

For each bootstrap iteration, six χ^2 tables (one for each metric) are generated. Each powerlifting score's χ^2 table can be used to judge how well the score ranks lifters with respect to a specified demographic variable. Perhaps the most important of these demographic variables is weight class. A χ^2 statistic for a powerlifting score based on weight class distribution would display counts of lifters within each weight class for each skill level, comparing it to hypothesized counts based on null proportions within each weight class for the full federation. The proportion of lifters within each weight class for the original sample of IPF lifters is displayed in Table 2. For each bootstrap sample, these proportions are calculated then used to obtain the hypothesized counts of lifters within each cross section of χ^2 tables.

χ^2 Goodness-of-Fit Statistic Calculation for a Powerlifting Score

$$\chi^2 = \sum_{All\ Cells} \frac{(Bootstrap\ Sample\ Count\ -\ Expected\ IPF\ Sample\ Count)^2}{Expected\ IPF\ Sample\ Count}$$

Expected IPF Sample Counts are the expected counts of lifters within a classification/skill level; These are calculated by taking the product of the classification's sample size and the proportion of lifters in the demographic level of interest within a full IPF bootstrap sample of size 59,047.

The goodness-of-fit statistic above can be calculated for each of six tables for each iteration of the bootstrap sample. The lower the test-statistic, the less biased that powerlifting score is for the full sample. Each metric has a test statistic for each iteration of the bootstrap.

This analysis utilizes 10,000 iterations of bootstrap resampling. Each iteration produces 6 χ^2 statistics (one for each of six powerlifting scores) for each of 3 demographic analyses (weight class, sex, and equipment). From these test statistics, a bootstrap distribution of test statistics can be assembled for each metric for each of 3 separate demographic analysis. Assembling bootstrap distributions for each metric provides an idea of how well a metric does in terms of distributing lifters by ranking. From these distributions, differences in χ^2 statistics can be taken to see which metric outperforms another. Particular contrasts of

Table 2: Proportion of All IPF Lifters by Weight Class (Original IPF Sample)

Weight Class (kg.)	43	47	52	57	59	63	66	72	74	83	84	84+	93	105	120	120+
Proportion of IPF Lifters	.002	.012	.047	.042	.042	.065	.048	.074	.093	.132	.050	.038	.141	.111	.064	.038

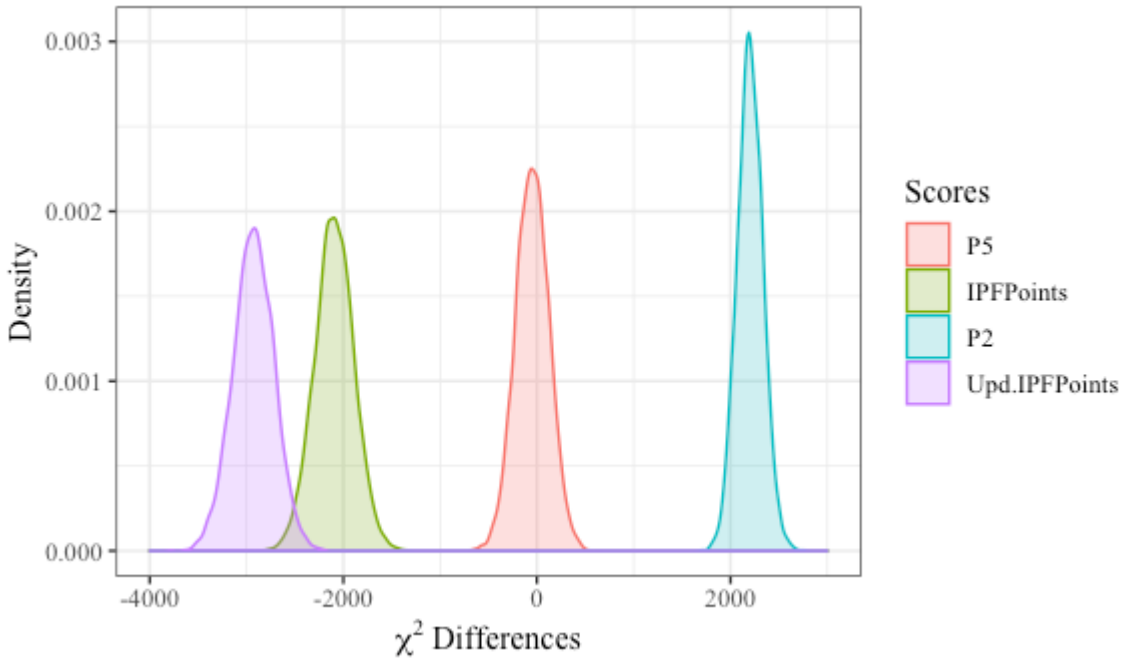
interest would be all metrics with Wilks. A positive difference in test statistics would indicate a larger test statistic for the reference/first metric, suggesting the subtracted/second metric matches the IPF sample better. To make interpretation easier, Wilks will be the reference metric. Thus, a positive difference in test statistics would suggest that the metric being compared to the Wilks Score more closely matches the IPF sample. Conversely, a negative difference would indicate that the reference/first metric, Wilks, matches the IPF sample better in that iteration of the bootstrap. Distributions of differences in test-statistics and quantile-based confidence intervals are used to demonstrate pairwise comparisons in the Results section.

4 Results

4.1 Bias by Weight Class

For each bootstrap sample, lifters were given a Wilks Score, Updated Wilks Score, P2, etc. All lifters were then ranked by each metric for each iteration of the bootstrap, then divided into classifications determined by quantiles (see Table 1). Within each classification, the number of lifters within each weight class was obtained. Each cell in a bootstrap sample's contingency χ^2 table corresponds to this count of lifters in a given weight class for all of the classification levels. After completing 10,000 bootstrap samples of IPF lifters, χ^2 statistics were obtained for each metric in each sample. Further, differences in test statistics between metrics were taken to make comparisons of interest. The original Wilks Score performance was compared to all other metrics given that Wilks has been the gold standard of power-lifting metrics for decades. As of 2019, the IPF's standard for best lifter was determined by IPF Points, replacing the Wilks Score as the competition standard. Thus, comparing IPF

Figure 3: Differences in χ^2 Goodness-of-Fit Statistics for Weight Class between Wilks Score and Other Scores



*Density for Upd. Wilks omitted due to large negative difference in χ^2 statistics

Points to Wilks and all other metrics remains pertinent for future IPF competitions. The relative performance of metrics can be visualized with density plots or histograms of the differences in χ^2 test statistics. Differences were calculated as the reference metric (Wilks) minus the new metric. Therefore, a positive difference in test statistics would suggest that the reference metric had higher χ^2 values than the new metric. This would mean the new metric performed better in bootstrap samples than the reference metric, suggesting that the new metric would be better at matching the observed IPF representation by weight class. Conversely, if differences were negative this would indicate that the reference metric is better than the new metric in terms of bootstrap sample performance.

In Figure 3, densities of differences in test statistics between the Wilks Score and other metrics are displayed for 10,000 bootstrap samples of all 59,407 IPF lifters. In χ^2 statistics for goodness-of-fit, the better the observed counts match expected counts in each cell of a table, the lower the test statistic that table will produce. Thus, the lower the test statistic,

the better the metric for that bootstrap sample. In this case, each table for a specific metric has counts for the skill level (classifications found in Table 1) by each weight class. To compare metrics, the difference between two metrics for the same bootstrap sample provides an indication of performance.

Each of the densities appears to be negative with the exception of one comparison. Wilks outperforms an Updated Wilks score greatly, with a massive negative difference in test statistics. This suggests that simply updating the Wilks formula is not the solution to providing better representation by weight class. The same idea of a fifth order polynomial seems to be an overfit compared to the original Wilks fifth order polynomial model developed from much less data (this is assumed given Open Powerlifting was not available when Wilks was last updated in 1997). Other models such as the Updated IPF Points and IPF Points are outperformed by Wilks, but by a much smaller margin. The P5 Score's distribution lies mostly above 0, suggesting comparable if not slightly better fit than the Wilks Score. The one model with entirely positive differences in test statistics from Wilks, suggesting a better fit for bootstrapped samples of IPF lifters was the P2. This model, like the model that produced the Wilks Score, has different coefficient estimates for each sex while also factoring in body mass. Where this model differs is in its adjustment for equipped or raw lifters, as it provides different estimates for each. Additionally, the P2 is much less complex than the Wilks model. Wilks uses polynomial terms of body weight up to the fifth power, resulting in more polynomial coefficient estimates using body weight. The P2 estimates a lifter's total using body weight to the first and second powers, so less coefficients are estimated for body weight. The performance of novel models compared to IPF Points follows suit of their performance against Wilks.

In the case of Weight Class, which has been made the primary focus of analyses of powerlifting statistics, P2 is by far the least biased. The quantile-based 95% confidence interval for difference in test statistics between Wilks and P2 is (1943.783, 2472.808) which is well above 0, suggesting there is compelling evidence that P2 is less biased than Wilks

from bootstrap sample to sample (Table 3). All other models' intervals for differences in bootstrap sample test statistics from Wilks were negative in value and large in magnitude or included 0 indicating Wilks provided the same or better fit for the same bootstrap IPF samples.

For most skill levels, the Wilks Score has minimal bias. However, for the two most skilled groups of lifters, the Wilks Score is moderately biased. Specifically, bias is observed in the top 5% of lifters by Wilks (Master level) and top 2.5% by Wilks (Elite level) (see supplemental materials).

In the top 2.5% performers by Wilks, there is a favorable, moderate bias toward the three lightest male weight classes (59, 66, 74). By Wilks Score, there was found to be a higher proportion of Elite lifters that fell into those weight classes than the proportion of all lifters in those weight classes for bootstrapped samples of the entire IPF. Conversely, Wilks is moderately biased against lifters in the 72, 84, 84+, 93, 105, and 120 kg weight classes (see supplementary materials).

The Wilks Score is less biased for Master level lifters than Elite lifters, but is still noticeably more biased than all other lifter skill levels. For Master lifters, Wilks slightly favors those in the 52, 59, and 66 kg weight classes while failing to compensate properly for those in the 63, 72, and 84 kg weight classes (see supplementary materials).

The best performing metric, the P2 is the least biased by far, with no evident signs of favoring lifters in any weight class by skill level.

4.2 Bias by Sex

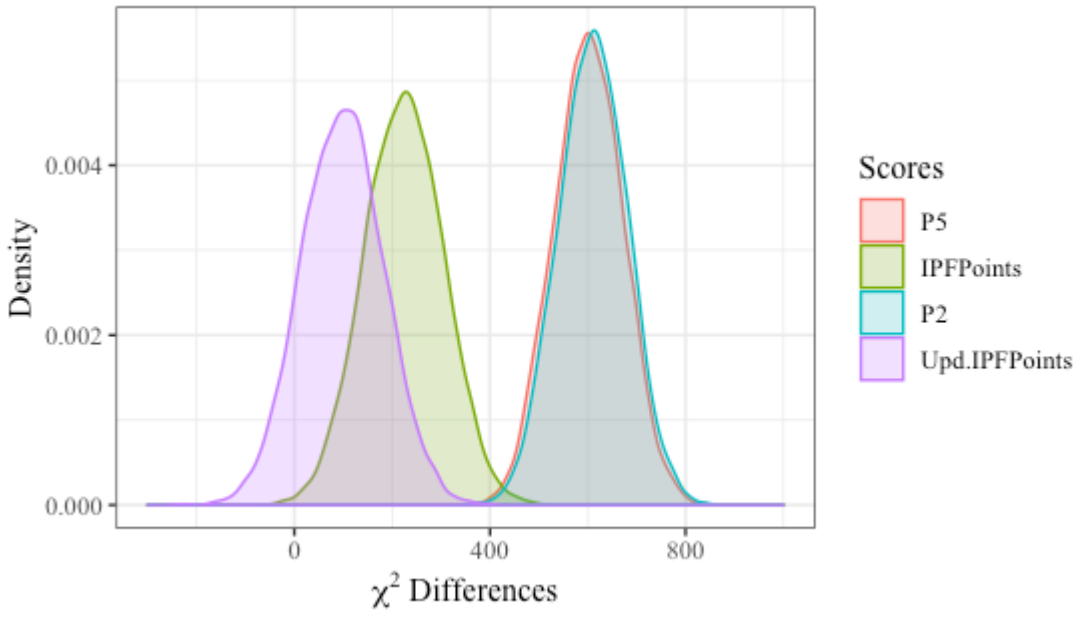
In the IPF data that has been filtered to include only each unique lifter's best total, of the 59,407 people, 66.5% were male and 33.5% were female. These would be designated as the hypothesized proportions for each sex in the original IPF sample but may differ based on different bootstrap samples of the IPF data. The idea is that within each sex and classification, the proportion of sex represented in each classification-based bootstrap

sample should match the proportion of males and females within each full IPF bootstrap sample. Variability in bootstrap samples of all IPF lifters is inevitable given the high volume of observations, so the null proportion will likely differ greatly from sample to sample. The density plots of differences in goodness-of-fit statistics on the basis of sex yield similar results compared to the same plots for weight class. By far, the worst performing metric in terms of bootstrap sample-to-sample bias was the Updated Wilks Score. Its differences in test statistics from Wilks and IPF Points were extremely high in negative magnitude, with its entire distribution below -43,000 for both pairwise comparisons. All other metrics perform very similarly in terms of biases. The symmetric distributions comparing Wilks and all other metrics except the updated Wilks are all entirely above 0 or mostly above 0, seen in Figure 4. On the basis of sex, this suggests that Wilks does not provide a good fit relative to most of the novel scores or IPF Points. In the distributions and confidence intervals for differences in χ^2 statistics for weight class, it was seen that the second-order polynomial P2 and the fifth-order polynomial P5 provided the best fit to the bootstrapped IPF data. Given the same performance in terms of sex, the use of a simpler P2 model may present itself as more appealing than a fifth-order polynomial like Wilks or the P5.

From Figure 4 the Wilks Score is more biased on the basis of Sex than all other metrics with the exception of an Updated Wilks Score. To support the visualizations, 95% confidence intervals for differences in goodness-of-fit statistics are displayed in Table 3. Intervals constructed from the subtraction of other metrics' test statistics from those of the Wilks Score are all positive, implying less bias from sample to sample in 10,000 iterations of bootstrapping the IPF lifter sample. Though P2 has the highest positive quantities in its interval for differences (474.659, 747.213), the P5 also outperforms Wilks noticeably with a similar interval (466.398, 737.552). Despite similar performance on the basis of Sex, simpler models are generally preferred if use is justified. The P2 estimates less coefficients and also is much more theoretically justifiable than a fifth-order polynomial.

With the exception of an Updated Wilks Score, all metrics seemed to perform the worst

Figure 4: Differences in χ^2 Goodness-of-Fit Statistics by Sex between Wilks Score and Other Scores



*Density for Upd. Wilks omitted due to large negative difference in χ^2 statistics

at the Elite level. The best performing model, P2, placed a slightly higher proportion of female lifters into the Elite category than was expected. This was also observed in the P5 model. Conversely, the Updated IPF Points and IPF Points Scores placed males in the Elite a little more frequently than expected. The Wilks Score performed worse than everything except the Updated Wilks Score, with a moderate bias in favor of male lifters at the elite level.

4.3 Bias by Equipment

In the sport of powerlifting, raw and equipped lifters compete in the same weight classes, but not against each other. For this reason, comparisons of raw and equipped lifters are often disregarded. Returning to the idea of the Wilks Score, the goal is often to know who the best lifter is regardless of weight class. Best lifter is often determined within boundaries of sex and equipment, so separate awards go to men's equipped, women's raw, etc. This separate

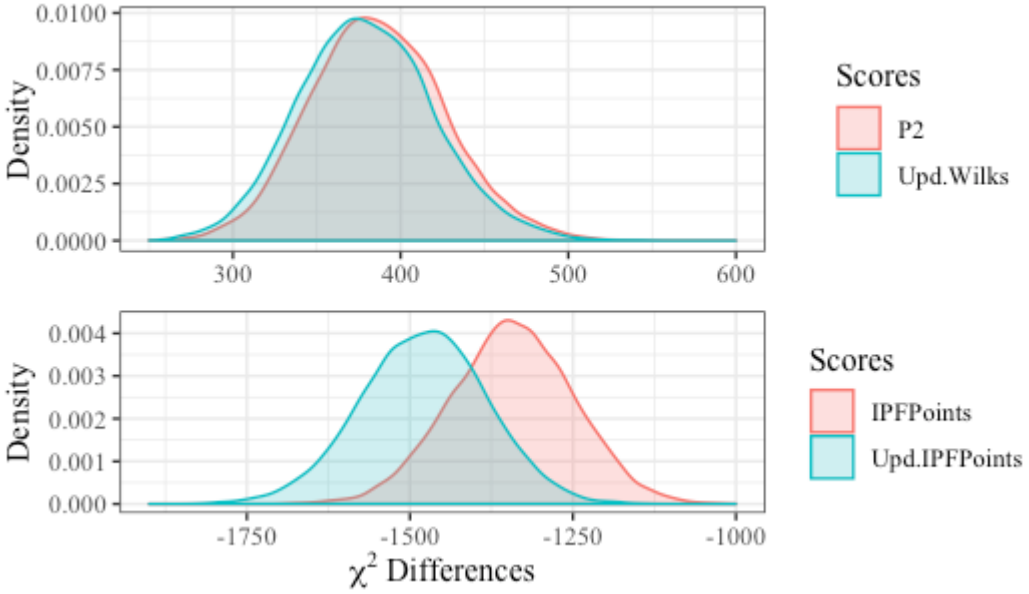
delegation of best lifter awards may solely be due to the Wilks Formula not adjusting for sex. In the same breath, it is important to note that IPF Points is the new standard for awarding best lifter at the highest level of competitions within the IPF. Given that IPF Points adjusts for sex and equipment choice, it may be useful for determining who is the most dominant lifter depending on its performance in goodness-of-fit comparisons with other scores.

An intuitive way to adjust for equipment and measure the effect of it is to include it in statistical models for total kilograms lifted. This would allow for the comparison of raw and equipped lifters. It is widely accepted that equipped lifting should yield a higher total compared to raw lifting for a given lifter, but the degree of strength increase is variable from person-to-person. Statistical models that include equipment, as well as its interactions with other predictors such as body weight and sex will properly adjust for differences between raw and equipped lifters.

The full IPF lifter sample is comprised of 73.8% raw lifters and 26.2% equipped lifters. Ideally, an unbiased metric should have the same distribution of raw and equipped lifters by each classification as the full IPF bootstrap sample. For example, the top 2.5% of lifters (elite classification) should be 60% raw and 40% equipped if a bootstrap sample of IPF lifters is split 60/40 in terms of raw and equipped, respectively. This should apply to all of the quantile-based samples. In Figure 5 below, the P5, a fifth-order polynomial regression adjusting for sex, body weight, and equipment use is by far the most biased by equipment status.

The best performance of the Wilks Score relative to the other metrics can be observed in the Equipment category. In Figure 5, only two metrics have densities above zero indicating less bias. Albeit a narrow margin, the P2 and Updated Wilks have density curves of differences in χ^2 statistics from Wilks in the low hundreds. Confidence intervals for differences in test statistics support the idea that P2 and Updated Wilks are the two least biased metrics evaluated. With 95% confidence, the difference in test statistics between Wilks and the P2 and Wilks and Updated Wilks, are between 311.723 and 469.097, and from 303.960

Figure 5: Differences in χ^2 Goodness-of-Fit Statistics for Equipment between Wilks Score and Other Scores



*Density for P5 omitted due to large negative difference in χ^2 statistics

to 462.781, respectively (Table 3). Similar to the evaluation of bias on the basis of Sex, a fifth-order polynomial provided comparable performance to the P2 by outperforming both Wilks and IPF Points. Still, the P2 is more simplistic and justifiable when compared to more complex models.

By equipment, the P5 Score is the most biased, followed by Updated IPF Points Score and IPF Points. Wilks, Updated Wilks, and P2 perform similarly, except Wilks is slightly more biased for Master and Class 5 level lifters. The Wilks Score places a higher proportion of Equipped lifters into the Master level than expected. For Class 5 lifters (individuals in between the 10th and 25th percentiles), Wilks gives a slight edge to classic/raw lifters.

5 Discussion

Visual comparisons of the density curves for differences in test statistics suggested the P2 model yielded the best results in terms of matching bootstrap IPF samples. The highest

positive difference from the Wilks Score and IPF Points in χ^2 test statistics indicates that P2 clearly outperforms both on the basis of weight class, sex, and equipment. While P2 is the least biased metric when evaluating by a lifter's sex and the equipment used in competition, the most impressive component is its performance in classifying lifters by weight class. P2 outperformed all other metrics on the basis of weight class by a wide margin. The performance of P2 on the basis of equipment and biological sex were matched by fifth-order polynomials, the Updated Wilks and P5, respectively. Inherently, it becomes easier for complex models such as fifth-order polynomials to correctly approximate distributions in which there are only two categories, such as biological sex and equipment. This is why similar performance is seen between P2 and fifth-order polynomials in terms of bias by sex and equipment.

Previous evaluations of bias in the sport of powerlifting have been limited in abilities to account for future variability in strength. The initial development of the Wilks Formula relied on a small set of data. IPF Points, while it was demonstrated to be the least biased proposal two years ago, was developed without any account of variability within the IPF sample. Powerlifting is continually growing with more competitors each year and records being continually broken. The methods previously employed by Marksteiner in his IPF Points Proposal [15], as well as the IPF in its evaluations [6], assume its data are the IPF population instead of a sample. This assumption requires future variability in the strength of powerlifters to be homogeneous for these methods to hold, or that its "population" is representative of the future of powerlifting. In contrast, the method of bootstrapping used in this analysis of bias purposely introduces more variability to the IPF sample of approximately 60,000 lifters to see if statistical models can account for future uncertainty. Additionally, differences in performance of IPF Points from 2017 to 2019 is likely to be due to the failure to account for variability. Further, the same specifications that were required for the IPF Points proposals two years ago for subsetting the IPF lifter sample were followed in the development in the models for P2, P5, and Updated IPF Points. Models in this analysis were trained on nearly

twice as many data points as Marksteiner's model trained on about 30,000 lifters [15, 3], which attests to the growth of the sport and reiterates that the basis of evaluation should include elements of variability. Accounting for variability was one of the considerations when evaluating how bootstrap samples would be evaluated. Particularly, the selection of quantiles for rankings by each metric was carefully considered. The IPF's most recent evaluation used quantiles in increments of 0.1 [6], makes distributional matching not as difficult as it should be because each quantile is the same size. Equal increments of quantiles do not truly answer the question that many figures in the powerlifting community have posed: Which score is fair for everyone and can also be used to select the best lifter? Evenly spread quantiles might answer the first part of the previously posed question, but it fails to account for the Elite classified lifters. The IPF evaluates the Elite lifters as the top 10% of lifters, which is incredibly generous. In the data used for this analysis, the 90th percentile (minimum value to be in the top10% of lifters) and 97.5th percentile (the USAPL Elite bottom cutoff) are separated by totals of 125 kg/275.51 lbs (685 kg compared to 810 kg). A difference that great in lifter totals suggests that the using the top 10% as the 'Elite' lifters may be insufficient for generalizing results to lifters with higher totals, and ultimately choosing a best lifter with IPF Points. Other authors such as Anna Khudayarov point out how IPF Points is great for explaining discrepancies in amateur lifters but fails to do the same when looking at the upper echelon of powerlifters so Wilks is still used outside of the IPF to choose the best lifter [12]. The P2 is able to be achieve both minimal bias amongst Elite lifters and amateur lifters across various demographics, providing a solution to problems posed by IPF Points and Wilks scoring, respectively. This analysis used the USAPL lifter classification quantiles (Table 1) which are not in uniform increments [18]. The USAPL is the U.S. IPF affiliate, the IPF's largest sub-federation, and is the only federation within the IPF with clearly documented lifter classification standards. Another reason for using the USAPL's guidelines is that conclusions from this analysis may be applied to federations outside of the IPF who may use the same classification standards despite slightly different weight classes.

One federation that has nearly all the same classification labels (Elite, Master, Class 1, etc.) is the American IPL affiliate, USPA [2]. Quantiles for the USPA's classification standards are ambiguous, but given the proximity of the USAPL and USPA in having shared lifters and inevitably intertwined histories, it is highly probable that one federation's standards came from the other's, making them the same today.

The methods of evaluating biases in this paper were limited to training models, selecting best models from Adjusted R^2, testing RMSE, and AIC, and evaluating their performance from non-parametric test statistics from 10,000 bootstrapped samples. This analysis' primary goal was to build statistical models with extractable coefficients to turn into novel scores, comparing them to existing model forms. The secondary goal was to have one of those models perform well in matching the IPF distribution across lifter demographic variables and all skill levels. Both of these were achieved. While the goals of this analysis were met, there are limitations to this style of analysis. To account for sample-to-sample variability in an explanatory model for strength, model-averaging across random samples of lifters could be an alternative to the use of 10-fold cross validation in this analysis. Model averaging for quantile regression models would be another way to model strength fairly for powerlifters if the method of analysis involves division of lifters into quantiles, similar to this analysis. Instead of a nonparametric approach such as comparing differences in χ^2 statistics, other practitioners can utilize parametric models that rely upon asymptotic approximations and continuity corrections introduced by authors such as Goodman [10], Agresti [1], and Jeong et. al [11]. Further analyses that build models to quantify strength in powerlifting should explore statistical models outside of polynomial regression, as there are limitations in model accuracy with this method despite its practicality when compared to more advanced techniques used in modern data science. Incorporating modern machine learning algorithms will likely be met with hesitation by the powerlifting community but would almost certainly provide scores that represent lifters fairly. For the P2 Score, it should be evaluated every two years, especially as the sport grows and more data are available. If fairness is of high

Table 3: 95% Quantile-Based Confidence Intervals for Differences in χ^2 Statistics

Variable	Score Comparison	(Lower Bound, Upper Bound)	Which is better?	Which is least biased?
Weight Class (kg.)	1. Wilks - P5	(-384.413, 288.814)	Inconclusive	P2
	2. Wilks - IPF Points	(-2495.625, -1695.404)	Wilks	
	3. Wilks - P2	(1943.783, 2472.808)	P2	
	4. Wilks - Updated IPF Points	(-3350.996, -2518.761)	Wilks	
	5. Wilks - Updated Wilks	(-42605.30, -41859.51)	Wilks	
Sex	1. Wilks - P5	(466.398, 737.552)	P5	P2
	2. Wilks - IPF Points	(68.327, 380.198)	IPF Points	
	3. Wilks - P2	(474.659, 747.213)	P2	
	4. Wilks - Updated IPF Points	(-63.894, 262.773)	Inconclusive	
	5. Wilks - Updated Wilks	(-44116.280, -43564.92)	Wilks	
Equipment	1. Wilks - P5	(-59023.130, -57771.740)	Wilks	P2
	2. Wilks - IPF Points	(-1530.834, -1162.302)	Wilks	
	3. Wilks - P2	(311.723, 469.097)	P2	
	4. Wilks - Updated IPF Points	(-1673.232, -1290.946)	Wilks	
	5. Wilks - Updated Wilks	(303.960, 462.781)	Updated Wilks	

concern to lifters, the every-other-year evaluation should be adhered to strictly, especially with what has been observed with IPF Points' drop in performance over two years. However, models such as the P2 that are evaluated with uncertainty in mind should stand the test of time more so than those which do not account for variability. To solve this recurring issue of picking a fair powerlifting score, some structure has to be assumed. If a federation is concerned with fairness, it should gather a panel of statisticians and powerlifting experts to create a framework for evaluating a score. Once this process for evaluation is created, scores can be submitted for evaluation with regular updates. Picking a fair powerlifting score requires more clarity and structure within the community. Until governing bodies and lifters are explicit about what is desired in a powerlifting score, the community must rely upon methods such as Robert Wilks' and those utilized in this analysis to rank and compare powerlifters.

References

[1] Alan Agresti. Testing marginal homogeneity for ordinal categorical variables. *Biometrics*, pages 505–510, 1983.

[2] United States Powerlifting Association. Classification standards. `https://www.uspa.net/classification_standards.html`, 2020.

[3] Per-Olof Astrand and Kaare Rodahl. *Textbook of work physiology New York*. NY: McGraw-Hill, 1986.

[4] George Casella and Roger L Berger. *Statistical inference*, volume 2. Duxbury Pacific Grove, CA, 2002.

[5] L. Fabien. Telephone interview with r. wilks., 2008.

[6] International Powerlifting Federation. Ipf formula: Why it was time for a new ipf formula. `https://www.powerlifting.sport/rulescodesinfo/ipf-formula.html`, 2019.

[7] International Powerlifting Federation. Technical rules book of the international powerlifting federation. international powerlifting federation. `https://www.powerlifting.sport/fileadmin/ipf/data/rules/technical-rules/english/IPF_Technical_Rules_Book_2020.pdf`, 2020.

[8] Pierre-Marc Ferland, Marc-Olivier Allard, and Alain-Steve Comtois. Efficiency of the wilks and ipf formulas at comparing maximal strength regardless of bodyweight through analysis of the open powerlifting database. *International Journal of Exercise Science*, 13(4):12, 2020.

[9] Andrew Gelman and Guido Imbens. Why high-order polynomials should not be used in regression discontinuity designs. *Journal of Business & Economic Statistics*, 37(3):447–456, 2019.

[10] Leo A Goodman. Association models and canonical correlation in the analysis of cross-classifications having ordered categories. *Journal of the American Statistical Association*, 76(374):320–334, 1981.

[11] Hyeong Jeong, Myoungshic Jhun, and Daehak Kim. Bootstrap tests for independence in two-way ordinal contingency tables. *Computational Statistics and Data Analysis*, 48:623–631, 03 2005.

[12] Anna Khudayarov. Revising the wilks scoring system for pro raw powerlifting. *arXiv preprint arXiv:1903.10694*, 2019.

[13] Max Kuhn. *caret: Classification and Regression Training*, 2020. R package version 6.0-86.

[14] Jake Lever, Martin Krzywinski, and Naomi Altman. Points of significance: model selection and overfitting. 2016.

[15] Joe Marksteiner. Ipf points - proposed replacement for wilks coefficients. `https://www.powerlifting.sport/fileadmin/ipf/data/ipf-formula/IPF_Points_Proposal.pdf`, 2019.

[16] Mary L McHugh. The chi-square test of independence. *Biochemia medica: Biochemia medica*, 23(2):143–149, 2013.

[17] OpenPowerlifting. Openpowerlifting data.

[18] USA Powerlifting. Classification standards. `https://www.usapowerlifting.com/wp-content/uploads/2014/01/Raw-Classifications-kg.pdf`, 2020.

[19] Gordon K Smyth. Polynomial approximation. *Wiley StatsRef: Statistics Reference Online*, 2014.

[20] PAUL M VANDERBURGH and ALAN M BATTERHAM. Validation of the wilks powerlifting formula. *Medicine & Science in Sports & Exercise*, 31(12):1869–1875, 1999.