Review of

Method Proposals to Calculate Best Lifter Scores (Relative Scores) in IPF Powerlifting Competitions

Submitted to the IPF to replace the Wilks coefficients

Reviewers: Dr. C. Maiwald[1] | Dr. T. Mayer[1,2]

[1] Chemnitz University of Technology, Department of Research Methodology and Data Analysis in Biomechanics
[2] TecStat Analytics, Werdau

Chemnitz, Werdau 29.10.2018

Contents

1 Proposals available for review and terminology used

A total of four proposals were available for review. Each proposal contains what we hereafter refer to as *method* or *model* to calculate relative scores in powerlifting. The methods and models will be dealt with in alphabetical order of the respective author's surname, and for simplicity will also be referred to by the author's surname hereafter:

- Unnamed method by Author 1, referred to as **Author 1**
- Author 2, Author 2.1 & Author 2.2: *Relative Strength Performance Model*, referred to as **Author 2**
- J. Marksteiner: *IPF Points*, referred to as **Marksteiner**
- Author 4: *Deciton Equivalent*, referred to as **Author 4**

The following abbreviations will be used in the remainder of the review

- MEN: men's category of lifters
- WMN: women's category of lifters
- PL: powerlifting, total score of three disciplines
- BP: bench press
- CL: classic, raw category of powerlifting
- EQ: equipped/single-ply category of powerlifting

2 Contents of this review

The review consists of two main sections.

1. Detailed discussion of the proposals and models (starting on page 6). Each proposed method is discussed separately. Proposals are discussed with respect to the scientific reasoning and theoretical background of the modeling approach (e. g. what model is chosen and why? How well is the model choice backed up by scientific reasoning?), how the chosen approaches differ from those of other proposals, the type of data used to develop the method, and its applicability.
2. Score comparisons across methods and performance levels starting on page 14. This section contains detailed numerical and graphical comparisons of the proposed methods in comparison to Wilks scores.

For readers who do not wish to examine statistical details or read the entire discussion, we provide a brief summary of the entire review on page 5.

3 Summary of the review

Not all of the submitted proposals included enough information to reproduce the scoring results presented in the papers. Author 1 is still unfinished, and presumably contains errors in the formulae we were unable to resolve. Therefore, we commented on Author 1, but did not include it in the numerical comparisons.

For each method that the reviewers were able to reproduce (Author 2, Marksteiner, Author 4), scoring results were created based on the datasets provided by Joe Marksteiner. Comparisons were made under two different perspectives:

1. The scoring results of the top lifters should mirror the population percentages of lifters across weight classes to ensure fairness in relative scoring. This is a purely distributional perspective, and a slightly weaker criterion than the one to follow.
2. The scoring should result in similar scores within a certain performance level. An ideal scoring method would result in e. g. the average scores of the best athletes in the lower weight categories being identical to the average scores of the best athletes in the middle and heavy bodyweight classes.

The consequence of both perspectives is that the likelihood to become best lifter is unbiased across weight classes. This would remedy most of the criticism brought up against the current scoring method using Wilks points.

To objectively evaluate the scoring performance of each method, we used both of the above perspectives. χ^2-statistics were used to measure distributional fairness, and a regression approach was used to determine similar average scores across the most important performance level of the top 10% of lifters. The regression approach was evaluated statistically and graphically. The combination of the two assessments allowed the reviewers to arrive at a comprehensive verdict that took both of the above perspectives into account.

In summary, we found that both Marksteiner and Author 2 are well worked out methods, both with advantages, drawbacks, and scientific foundation. **When given the choice, Marksteiner scores can be labeled as the fairer system when all subdisciplines and performance levels are taken equally into account (regardless of weighting). When focusing on elite & top 20 % lifters and weighting both components of our analysis (χ^2 & Loess scores) equally, Marksteiner again performs better than Author 2. When weighting χ^2 scores half in comparison to Loess scores, Author 2 performs better for elite, top 20, and top 30 % of lifters.** However, the reviewers want to emphasize they currently see no reason to apply such a χ^2 attenuation.

All three evaluated methods and models have their advantages and drawbacks. We were unable to include the Author 1 method in numerical comparisons. It contains a well laid out scientific basis, a promising modeling approach, but also an additional factor of *age adjustment*. The reviewers speculate that such an age adjustment could result in even better model fits and increased fairness compared to the other three methods, which currently operate without age adjustment. However, the Author 1 method is still unfinished, and relative scoring to bodyweight *and* age may introduce an undesired level of complexity. Furthermore, the acceptance of age adjustment among the IPF and powerlifting community is still unclear at this time, and may be an issue for future improvements of relative scoring methods in powerlifting.

4 Detailed discussion of the proposals and models

The four proposals differ greatly in the amount and detail of given information. In part, this resulted in difficulties in reconstructing the mathematical and analytical operations performed by the authors. These difficulties will be addressed in the specific sections of the proposed methods.

The purpose of the four proposals is to provide a new method for calculating scores for the best lifter competition in IPF powerlifting. All proposals rely on methodologies that model relative strength (relative performance) as a function of body weight. Author 1 includes adjustment for age, but since the document is unfinished, the final formulae are not given and cannot be evaluated.
Author 2 also considers age as a factor, but does not include specific coefficients to be included in any evaluation of the developed formulae. We point out that this review will *not* evaluate age standardization procedures. It is unclear at this point in time whether standardization to age is easily applicable and – more importantly – even supported by IPF strategies or the majority of the IPF staff who are responsible for implementing a new best lifter methodology.

As a function of lifters' bodyweights, Marksteiner proposes details to calculate best lifter in all three subdisciplines plus PL, split into MEN/WMN and EQ/CL. Author 2 proposes a methodology which is also divided into MEN/WMN and EQ/CL, but only includes PL and BP, no deadlift or squat. Author 1 does not account for any subdisciplines, but separates according to MEN/WMN and EQ/CL. Author 4 only includes the MEN/WMN division in PL without differentiation between EQ/CL or any subdiscipline.

The proposals share a common point of origin: the deficiency of the currently-used Wilks method to calculate relative scores for the best lifter competition. Specifically, several different aspects of the Wilks methodology are pointed out by the authors, which they intend to remedy with their proposed methods:

1. **Unfairness:** Overrepresentation of certain weight categories in best lifter results. This bias is argued to result in the heaviest and lightest lifters being more likely to win best lifter competitions.
2. **Outdated / Inconsistent:** It is argued that Wilks methodology was developed based on outdated lifting scores of the mid-1990s. The current lifter population has evolved to include higher body weights and lifting performances. Thus, the coefficients need to be updated at best.
3. **Only includes PL:** Wilks methodology was developed from data of equipped lifting, and may not provide adequate results for classic/raw lifting.
4. **Inaccuracy / Regression Bias / Lack of theoretical foundation for the polynomial regression:** Polynomial regression used by Wilks can result in uncontrollable effects outside the functional domain of lifter performances used for model creation. This means that if the lifter populations shift with respect to body weight and performance, the polynomial regression may result in unreasonable results when applied to new or outlier data.
5. **Over-complexity (Author 2):** 5th order polynomials may result in overfitting and over-complexity. Simpler models may provide a better-suited approach and may eliminate some of the effects described in the previous point.

4.1 Author 1: (Unnamed proposal)

4.1.1 Scientific foundation and rationale

Author 1 argues that the resulting coefficients of the Wilks polynomial regression are unreasonable, especially outside the range of the body weights used by Wilks for model creation. In addition, Wilks scores on contemporary datasets are not distributed evenly across weight classes, with heaviest and lightest classes being overrepresented in higher Wilks scorings.

χ^2-tests are used to demonstrate the discrepancies in distributions of Wilks scores, as well as a means to demonstrate the remedy suggested by the Author 1 methodology. However, only dichotomous categories are used with thresholds of 425 and 525. In contrast to Marksteiner, this approach to "fairness" employs weaker criteria, and still allows for substantial difference in score variance across weight classes.

Author 1 employs gaussian regression to model relative performance as a function of bodyweight, using a squared exponential, adopted from the Sinclair coefficient used in Olympic weightlifting. In contrast to Marksteiner and Author 2, Author 1's approach does not assume an empirical or analytical model a priori. Gaussian regression is usually used in determining properties of systems with unknown underlying functions and mechanisms, as well as in analytical problems of interpolation and smoothing. Since the relationship between lifters' body weights and relative performance can be argued to include many unknowns and stochastic processes, Author 1's approach is reasonable. It effectively omits the model selection problem outlined by Author 2, but at the same time avoids the drawbacks of polynomial regression mentioned by Author 2 and the author himself.

Author 1 depicts more evenly distributed relative scores, and also provides evidence for increased fairness by non-significant χ^2-test results. Since we cannot reproduce the scoring (formulae issue), we cannot comment on the validity of this claim.

4.1.2 Criticism

- The proposal is incomplete and unfinished.
- Using gaussian regression is reasonable, and may eliminate the drawbacks associated with polynomial regressions, and allows more flexibility than some of the empirical distributions.
- Age adjustment is an interesting aspect. It is evident from Author 1's work (similar analyses also included in Author 2's appendix!) that age plays a major role in modeling the relationship between lifters physical constitution and performance. However, there may be issues with acceptance and the complexity of calculations. First, age adjustment will control for age and apply corrections, especially for older athletes. Their likelihood of becoming best lifter increases dramatically, possibly leading to a podium of the oldest people in the considered age categories. It is unclear at this point whether such an outcome is politically desired and accepted among athletes.
- Erroneous formulae are given.
  In the proposal, the formula is written as: $10^{a(log10((b/d)+c)^2-a(log10((b/x))+c)^2}$
  There is either a bracket missing or surplus in the exponent of the formula.
- Testing the superiority of the proposed method using a dichotomous outcome (above or below 525 in EQ, above or below 425 in CL) disregards scoring validity, and only assesses proportions of lifters belonging to arbitrarily defined scores. Whether lifters who are significantly above 425 points score similarly is not assessed by such a procedure, but can be considered the most important aspect in method assessment.

- Why is EQ and CL combined in one model? Mechanisms and causalities for achieving higher performance can be different between EQ and CL, and may also interact with bodyweight. Using only one model that does not properly fit appears unreasonable to the reviewers.

4.2 Author 2, Author 2.1 & Author 2.2: Relative Strength Performance Model

4.2.1 Scientific foundation and rationale

The authors use a robust regression method to derive what they call an asymptotically balanced model. Initially, they fit their model in direct comparison with a fifth order polynomial (Wilks) to a large dataset, which results in identical statistical properties as the currently used polynomial fit. However, Author 2 (similar: Author 4) argue that fitting the model to an average athlete is considered problematic, since the true dependency of strength and bodyweight will be masked by large numbers of athletes not performing at an optimum level.

Such an approach is justified for an analytical model, hence Author 2 et al. use a non-probabilistic, stratified sample of elite lifters (within 15% of world record for each weight class). Their model fits result in an adjustment power efficiency factor, which is used to determine the relative score.

Large parts of the work contain the scientific foundation for *why* the authors chose the approach of an analytical model. However, the actual part of reasoning how the model formula was developed, is currently not available in the proposal because of a pending publication procedure in a scientific journal. We thus cannot comment on one of the most fundamental parts of this proposal: the reasoning for why the analytical model was chosen in its current form.

Overall, Author 2 et al. present a detailed and well-founded work, the methods used and the structured approach suggest a deep understanding of the subject matter. However, some assumptions and parts of the methodology are critical or potentially ineffective in developing a new relative index.

Furthermore, the authors did not provide information on suggested update intervals for the model coefficients.

4.2.2 Criticism

The work of Author 2 et al. is tremendously comprehensive and detailed. However, some of the authors' claims appear questionable to the reviewers.

- The authors state that previous approaches of model selection had no theoretical basis, since the models for curve fitting would be selected exclusively on the basis of the 'best-fit' criterion. Based on this statement, the authors consider it necessary to develop an analytical model. These statements disregard the epistemological dispute between induction and deduction in model creation, and that all of the modeling approaches of physiological processes – especially in strength output of athletes – have to deal with a tremendously complex issue that cannot be fully understood with today's knowledge. Empirical models have well-founded theoretical bases – they just differ from the theory applied in analytical models. Whether an analytical or an empirical model is *better* in a particular context depends on many variables. A general superiority of analytical models over empirical models cannot be ascertained a priori. Hence, the reviewers do not agree with the authors' claim of the necessity to develop an analytical model.
- Furthermore, the found/developed analytical model cannot be evaluated in-depth, because it is part of a pending scientific publication by the authors. Consequently, the authors' following statement "*Thus, relying on analogies with the biological patterns of metabolic processes in the body; the empirical laws of the practice of powerlifting; the most fundamental principles of conservation of matter and energy, we can assert that the expression obtained has the most satisfactory theoretical justification, at least in*

comparison with those approaches that are known to us." is not satisfying. For instance, the determinants of lifter performance could be quite different under certain circumstances, especially for EQ and CL. Currently, there is no rationale given by the authors why the same analytical model can be applied validly to both disciplines.

- Selecting quota stratified samples may be advantageous in determining analytical models, but the resulting small sample sizes make model fitting more susceptible to outliers. It is questionable whether such a procedure will produce fair rankings for a larger population. E.g. for fitting a model to men's classic bench press, only two data points were available for model fitting in the range of 120 kg to approximately 150 kg. The population, however, contains a large proportion of athletes in this weight class (see Figure 1).
- The relative points calculated by the authors are linearly scaled up/down from the world record level of the fitted curve. This approach does not take into account that the standard deviation in the total lifting population increases with the weight classes. Hence, some athletes might be systematically favored or disadvantaged.

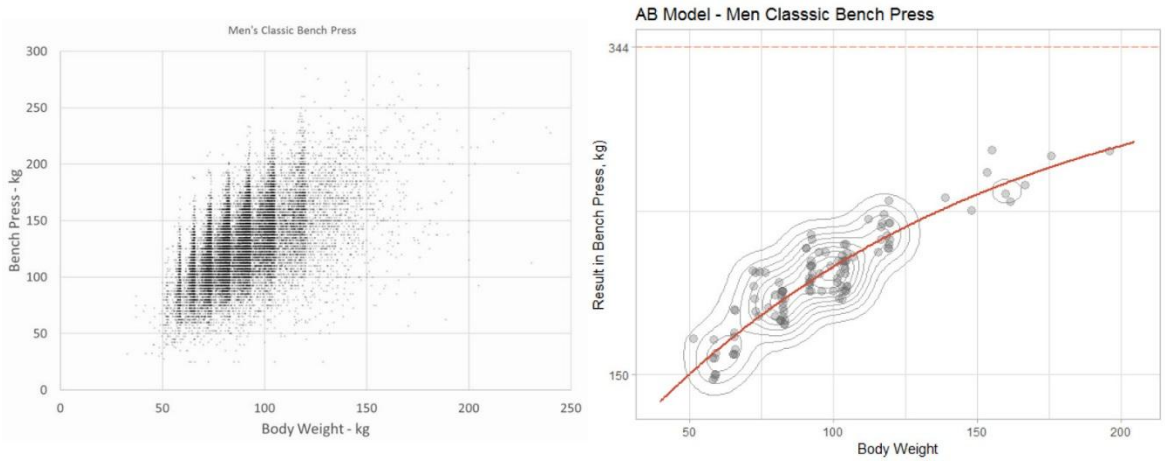

Figure 1: Comparison of available data points in the population and data points used for modeling. Figures taken from the Marksteiner and Author 2 proposals.

4.3 Marksteiner: IPF Points - Proposed Replacement for Wilks Coefficients

4.3.1 Scientific foundation and rationale

The method proposes to model lifter performance as a lognormal function of bodyweight. Lognormal distributions are often found in populations of biological systems, especially in scaling phenomena (e. g. length of limbs, bodyweights). However, two assumptions are elementary in this context:

1. The body weight of the lifting population must be approximately lognormally distributed.
2. The lifter performances in the different weight classes must be approximately normally distributed.

The author has examined the approximate lognormal distribution of body weight in the total lifting populations of men and women (see Appendix 5). However, the distributions of body weights in the populations of the (sub)disciplines, which had been fitted with the lognormal function, should have been checked, but either were not or were not presented.

After fitting the samples of the (sub)disciplines, Marksteiner uses an unconventional, but statistically correct approach to model the varying standard deviation across weight classes. This procedure is used to include varying standard deviations across the weight classes, which is essential in calculating correct standardized percentage rankings or deviations from the mean intended to be fair across weight classes.

The approach chosen by Marksteiner is reasonable, and is focused on providing a solution to the distribution problem of best lifters and their predominance in heavy and superheavy weight classes. The modeling approach is empirical, not biologically analytical like Author 2. Marksteiner does not attempt to model the physiological relationship between bodyweight and lifter performance based on selected data, but rather the empirical law which is accessible by observing the total population and their average performance. This approach rests on the assumption that all weight classes are populated with the same proportion of athletes, e.g. a similar distribution of athletes (ratio between elite and poor performers) can be found for each weight class. Using data sets with over 20.000 individual best performances across several years provides a reasonable basis for this kind of approach.

The author suggests updating the coefficient matrix to calculate mean performance and standard deviation every 4 years.

4.3.2 Criticism

- Based on Table 6 shown in Appendix 5, Marksteiner tries to demonstrate that the lifter performances within the weight classes in the (sub)disciplines are approximately normally distributed. In this point, his approach is not convincing, because the case numbers shown in the table are not consistent with the data sets used. Furthermore, the name of the correlation coefficient (Rho) indicates that a rank correlation method (Spearman) was used to check the assumption. The use of this method is not appropriate in this context and must be described as extremely questionable. Established standard methods here are Shapiro-Wilk tests (for normal distribution) and Q-Q plots (lognormal and normal distribution). Especially Q-Q plots would give the reader significantly more information for both distributions than e.g. Table 6 (Appendix 5). In summary, it remains unclear whether the two conditions mentioned above are really fulfilled and thus the theoretical assumptions for modeling with lognormal fits are

given. Marksteiner neither compares different models nor specifies the goodness of his fits, which would be correct if the 2 theoretical assumptions were fulfilled.

- The author has compared the results of his method to those of Wilks using correlations. The results are fully listed in the proposal for all (sub)disciplines (Table 7, Appendix 8). The comparison of scores by correlations alone may lead to misinterpretations. At this point, further comparisons using graphical methods would have been desirable to be able to better judge the differences/characteristics of both methods in direct comparison.
- Marksteiner's score can be specified in points and as a percentile value. The author leaves open which score should ultimately be used, but suggests that percentile scores are very understandable and easy to interpret for athletes & coaches. The reviewers do not share the view of the author in that a proportionate distribution of relative points (best lifters) across weight classes, compared to the distribution in the total lifter population, is a good criterion to ensure fair scoring. From our perspective, it is one side effect, but not a necessity. It cannot be guaranteed that all weight classes are equally populated with good and bad lifters (possible sampling bias).
- The work of Marksteiner contains several seemingly small errors, which significantly hinder the understanding of the described procedures for the reader and create inconsistencies within the proposal. For example, formulae on pages 6 and 19 are not identical, figures are incorrectly numbered, and correlations are not reported in a consistent manner.

4.4 Author 4: The Deciton Equivalent

4.4.1 Scientific foundation and rationale

Author 4's idea to design the relative score so that the score models the lifting performance of an athlete with a hypothetical body weight of 100 kg is tangible and quite interesting. In male athletes. a body weight of 100 kg is clearly within the range of “middle” weight classes and is still located quite centrally in the log normal distribution of body weights. It thus reflects the notion of comparing results to one of an “average” athlete of “typical” or “average” bodyweight.

4.4.2 Criticism

For the female athletes, a body weight of 100 kg is quite far from the +84 kg threshold of the open class and is oriented towards the right end of the log normal distribution. It is therefore highly unlikely that women will identify with this index to the same extent as men can. As a result, this index could encounter acceptance problems with female athletes.

In addition, the methods used in Author 4's proposal were neither substantiated nor backed by scientific theories. The following points are particularly critical:

- No theoretical reasoning for the model used
- Use of different models for men and women, without explanation (polynomial 6th order vs. 5th order polynomial)
- No information about the goodness of fit
- No substantiated rationale for the selection and composition of the fitting sample
- No separate models in (sub) disciplines, no discussion why this was not done
- Curve fitting with polynomials despite actually known problems, such as overfitting, over-parametrization etc.
- No meaningful evaluation of the developed score or comparison with the Wilks Score (solely presentation of individual cases)

In summary, it should be noted that theoretical considerations for this proposal seem to have played a less important role (in comparison to the other proposals) or were simply not described. Hence the proposal does not meet the same scientific standards as the other three. Author 4 was still included in the following comparison, since his formulae are worked out and, to the knowledge of the reviewers, correctly communicated in the proposal.

5 Score comparisons across methods

5.1 Methodology used in this section

We use the data provided by Marksteiner for illustration purposes and to resemble an entire population of lifters (given subsets PL/BP, MEN/WMN, and CL/EQ). In all of the following graphs, color coding will be as follows: Wilks (red), Author 2 (blue), Marksteiner (green), and Author 4 (purple). Since Author 1 did not allow for score calculation, it will be omitted from further analysis.

5.1.1 Model Fit Plots

To evaluate method characteristics and performance, we first plot their average prediction against bodyweight, and compare all applicable/available methods in one plot with respect to their predicted average of the entire population. Color indicates the respective methods. Plotting performance versus bodyweight with added model fits results in a graph that is found in nearly all proposals:

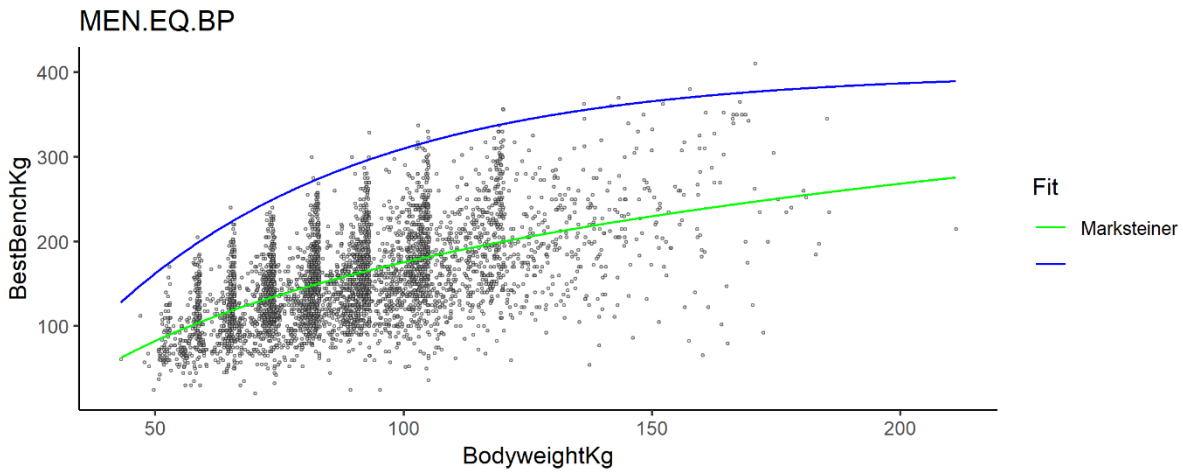

Figure 2: Model comparison of Author 2 and Marksteiner using the data of men's equipped bench press (n=4294). The differences in modeling philosophy are clearly visible.

From this type of graph, one can inspect the differences in the model selection with respect to the modeling philosophy, and how well the respective model fits the data of a population of lifters, to which it will be ultimately applied. Based on these fits, the relative scorings for each methodology are calculated according to the descriptions given in the proposals. Models and calculations were implemented in the statistical software R. Script code used to calculate relative scores is given on page 46, so authors can check the correct implementation of their methods.

Not all methods result in fits that can be plotted in a meaningful way. Note that Author 4 only predicts scores for PL. Applying the prediction method to BP data results in curves way above the data points, distorting the graphs and limiting their interpretability. Author 4 will thus be omitted from BP model fit plots.

5.1.2 Relative Scoring Distribution

To help visualize the impact of the methodologies on relative scoring distributions, we plot bar charts with relative frequencies of all scoring methodologies against weight classes for percentile groupings of 10%, beginning at the top 10% of lifters (P100) down to the weakest 10% of performances (P10).

An example plot is given in Figure 3 for men's equipped bench press in the P100 performance band (top 10 % of lifter scores). Based on the distribution of athletes across weight classes in the entire population (black bars), each scoring method introduces a distribution of P100-scorers across weight classes. Ideally, the distributions of the scoring methods match the population distribution across weight classes. Figure 3 indicates that e. g. Wilks scoring (red bars) results in the most extreme distribution bias among the methods, favoring heavier competitors and leading to their overrepresentation in the distribution of P100 scorers. This is a common claim made against the Wilks scores, which can be backed up in an objective manner using this type of data visualization.

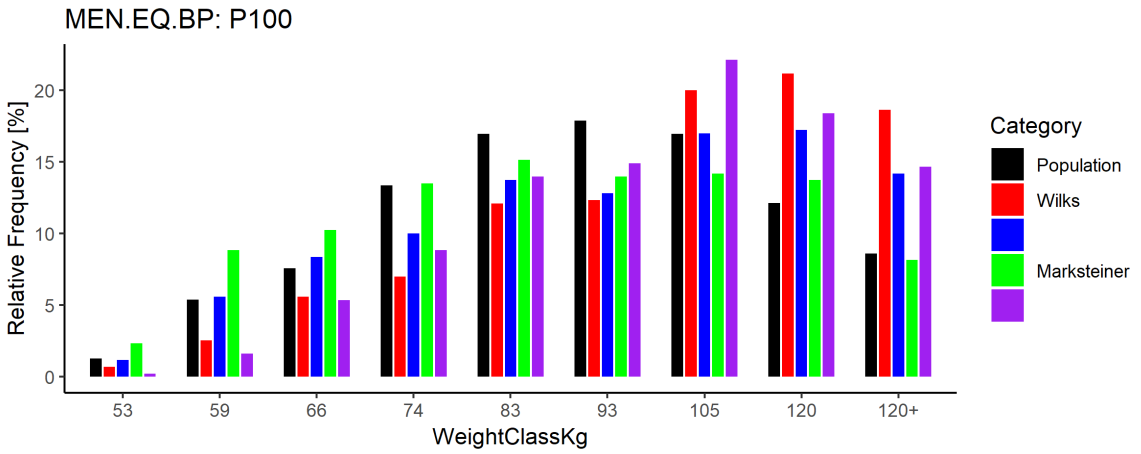

Figure 3: Relative scoring distribution (relative frequencies) for the reviewed scoring methods across weight classes. Almost all methods lead to significant overrepresentation of 105+ kg athletes among the top 10% (P100) of relative scores.

We use the χ^2-statistic to provide an objective measure of distributional proportion within each performance band, across all weight classes. χ^2 is a measure of how much the expected counts for the weight classes (determined by the population counts) are represented by the observed counts of the athletes in the specific performance band across weight classes. Since we do not make inferences to a population, we omit the commonly performed significance testing with this statistic. Although its numerical value does not allow for an intuitive interpretation as such, it can be directly compared across methodologies, with ***smaller* χ^2-statistics representing *less discrepancy* between expected and observed frequencies across the weight classes**. In Figure 3, the scores of Marksteiner distribute closest to the distribution of athletes in the population (black bars), hence the method would score the smallest χ^2-statistic among the methods.

The χ^2-statistics for each method are then summed across all performance bands, leading to a cumulated distribution bias score. Section Comparison of the scoring methodologies (page 18) will depict the χ^2-results, and contains further descriptions of how these scores contribute to choosing a best-suited method for best lifter scoring.

χ^2-statistics only check distributions, not scoring levels. That is why it only addresses one aspect of scoring fairness: under the assumption of a sufficiently populated sample and the basic principles of performance generation remaining constant across the range of weight classes, it may represent scoring fairness. However, by definition it cannot be used as a measure of scoring validity!

5.1.3 LOESS-Plots

In contrast to the *distribution effect* of methods depicted in bar charts, we need to add information on how much the methods affect *average scores across weight classes* in each performance band. This is to some extent independent of how proportionally athletes are allocated to the performance bands by the different methods. It rather shows how comparable the scores of the performance levels are across weight classes, within the method itself. Assuming that large samples of data contain equally performing athletes from all weight categories, average relative scores within a weight class should be the same across all weight classes. To check for this, we employ novel graphical methods and statistics to objectively quantify this feature of the scores and depict them in LOESS-plots.

The LOESS-plots used in the following sections consist of several layers of data. Figure 4: LOESS-plot with LOESS-fits for all performance bands (P10 to P100, red lines) in men's equipped bench press. First, we plot the relative score against bodyweight for each method in a separate plot. We then establish 10 performance bands based on the percentile groupings of relative scoring across the entire population, ranging from bottom 10 % (P10) to top 100 % (P100) in steps of 10 % each. Performance bands are visually indicated by alternating intensities of light and dark gray in the underlying data point cloud.

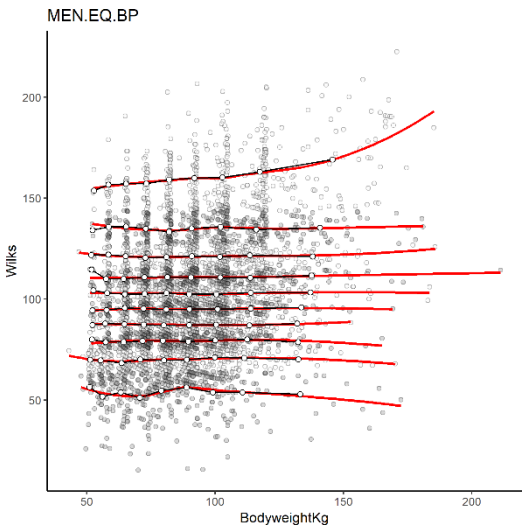

Figure 4: LOESS-plot with LOESS-fits for all performance bands (P10 to P100, red lines) in men's equipped bench press.

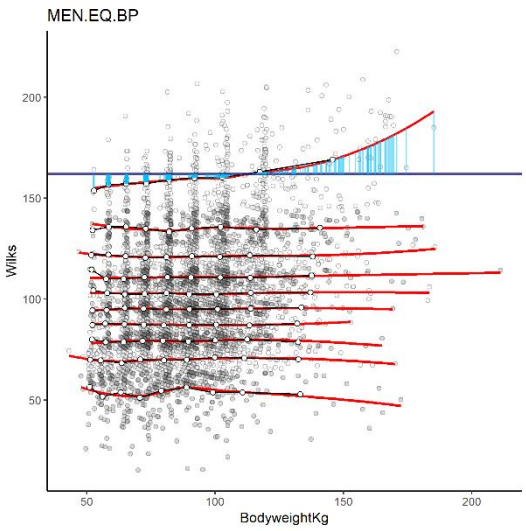

Figure 5: LOESS-plot with added P100 mean (dark blue line) and residual error of the LOESS-fit against the mean (light blue lines).

In the next step, layers of dots and lines are superimposed on the scatterplot. Within each performance band, thin black lines connect the means (white dots) across weight classes. These means are calculated for each subset of weight class and performance band. Solid-colored lines represent LOESS[1]-fits of average performance within the performance band across bodyweights.

Ideally, dots, thin black lines, and colored LOESS-fits would line up in a straight and *level* (!) line, much like for most of the middle performance bands in Figure 4. However, in the case of men's equipped bench press we observe a substantially increasing average Wilks score in the top 10%

[1] LOESS-fit is a type of **lo**cal regr**ess**ion that is used to fit models to data for which no suitable *overall* model is known. It fits linear or quadratic polynomials to local subsets of data and can accommodate data distributions that do not comply with many other modeling assumptions (e. g. homogeneity of variance or certain error distributions).

of lifters as their body weight increases. This reflects some of the criticism brought up against Wilks scoring, namely favoring heavier lifters over lighter lifters in some of the competitions. While the means for each weight class tend to tell that story to some extent, the LOESS-fit is more suitable to inspect the scoring behavior of the method across weight classes within a certain performance band.

To quantify the deviation of the LOESS-fit from an ideal horizontal average scoring result within each performance band, we calculate mean squared deviations of the LOESS prediction versus the horizontal line representing the mean of the LOESS-fitted values (see Figure 5). Since this deviation is dependent on the range of values of the specific scoring method, we normalize these deviations to total scoring range, and calculate percentage values. This is done for each performance band, and a final sum of the mean squared errors is calculated. It is given in its unweighted form in tables below LOESS-plots, and summarized in table 9 on pages 41 & 42. **The less residual error a method creates, the better the method performs in balancing the average scoring across weight classes.**

5.1.4 Effect on best lifter rankings

We *did not* calculate effects of the methods on the best lifter rankings for recent IPF events, as some of the authors did in their proposals. Such an approach may be informative to powerlifting experts, but does not contain any information which enables us to evaluate the quality of the method itself. The impact of the methodology on actual rankings is a pure consequence, and not the origin for determining validity or applicability. Validity and applicability are driven by the criteria mentioned above.

5.2 Comparison of the scoring methodologies

5.2.1 Men's classic bench press (MEN.CL.BP)

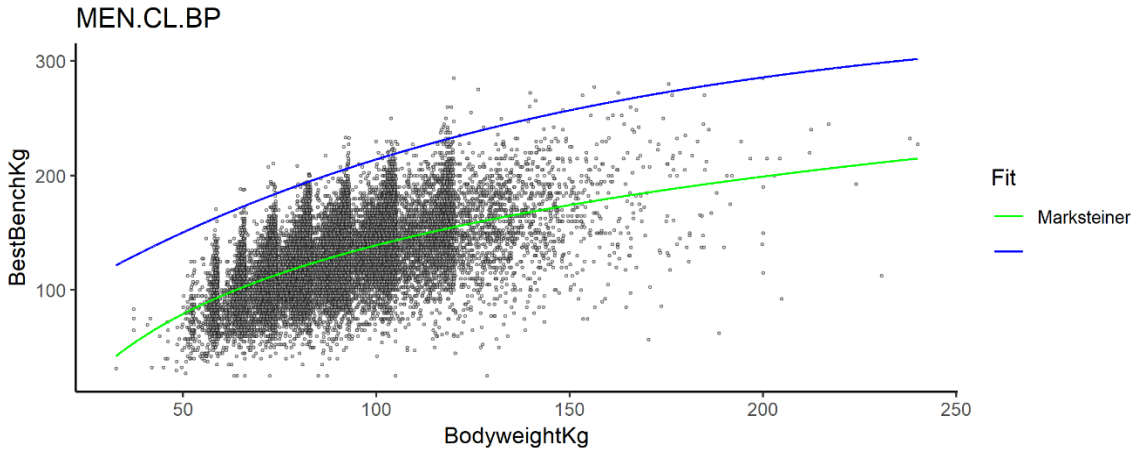

Figure 6: Model fits for men's classic bench press

Table 1: Summary statistics for men's classic bench press

	Wilks	Author 2	Marksteiner	Author 4
Distribution χ^2 sum	829.96483	1014.02677	280.71897	772.38166
LOESS residual sum	2.06503	1.91810	1.70323	1.83663

Marksteiner's scoring represents the shape of the underlying population best for the distribution of relative scores. LOESS residual sums indicate that Marksteiner scores are most level across all performance bands (see table 1 & figure 8).

Figure 7 on the next page depicts LOESS-plots with relative scoring across weight classes and performance bands in MEN CL BP.

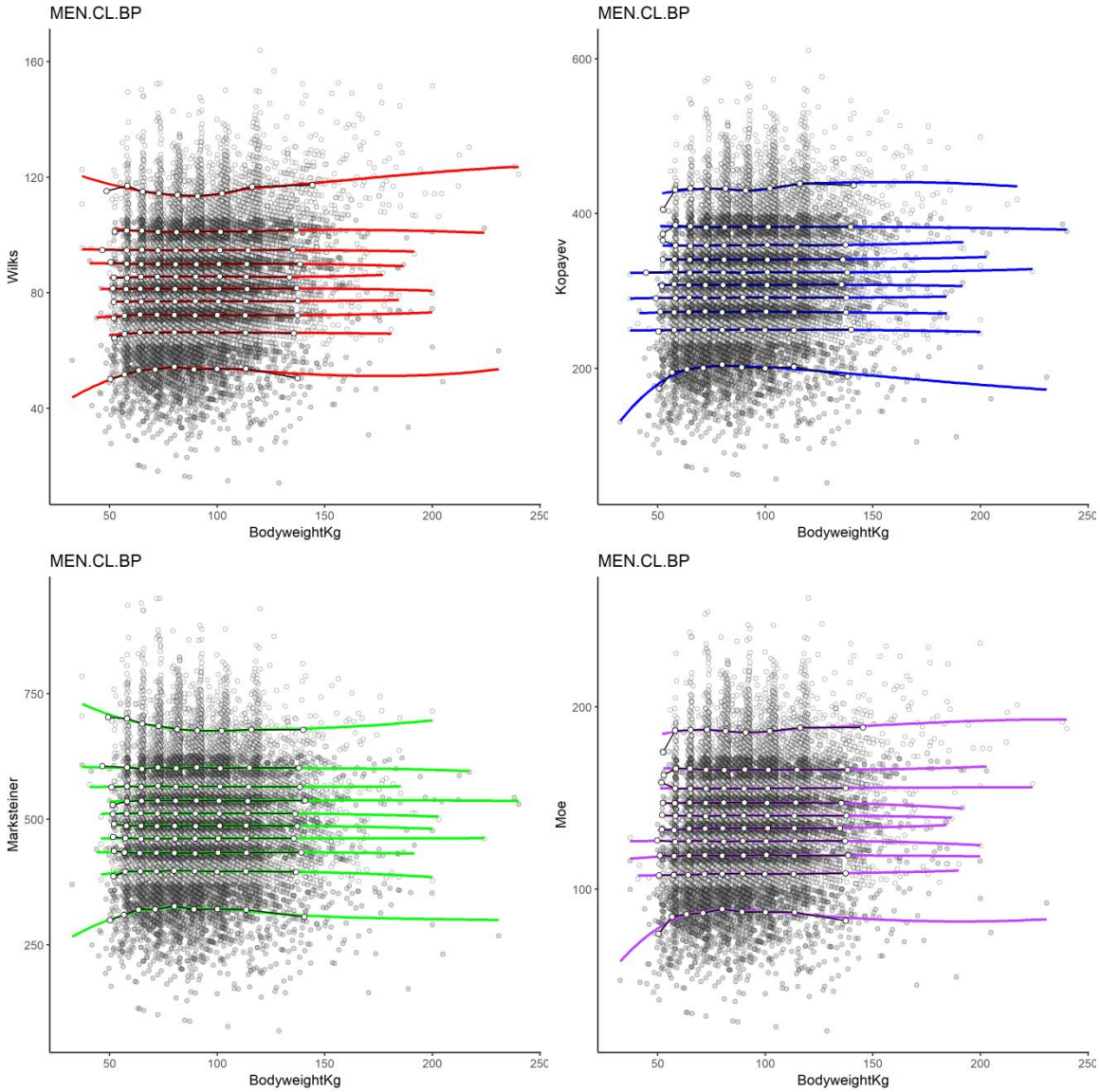

Figure 7: LOESS-plot for men's classic bench press

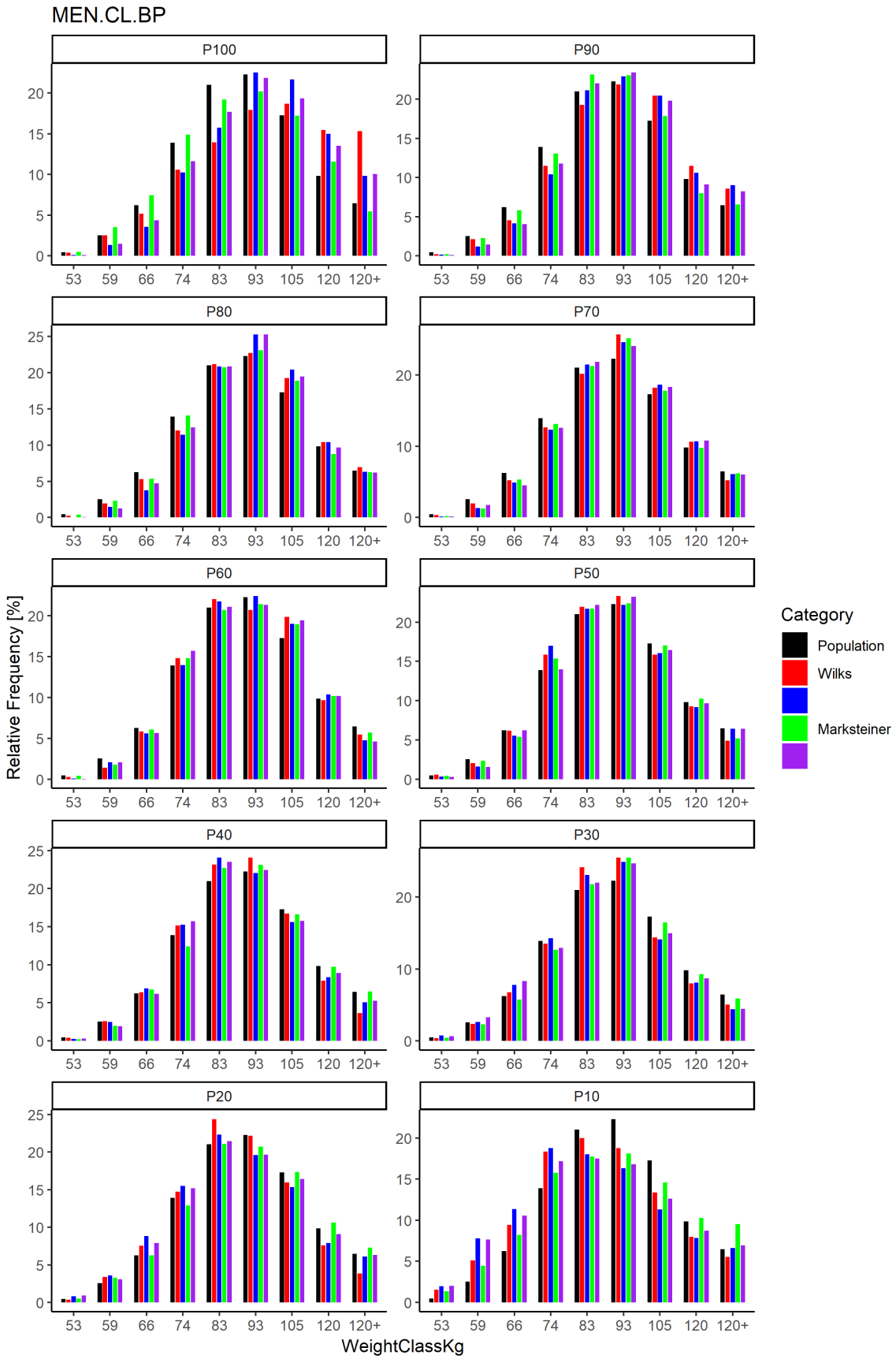

Figure 8: Distribution of relative scores across performance bands for men's classic bench press

5.2.2 Men's classic powerlifting (MEN.CL.PL)

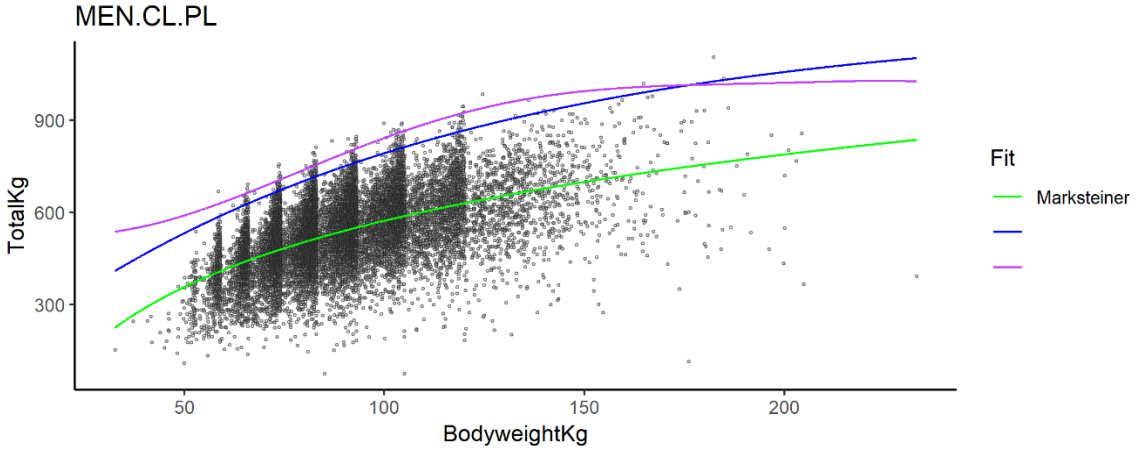

Figure 9: Model fits for men's classic powerlifting.

Table 2: Summary statistics for men's classic powerlifting

	Wilks	Author 2	Marksteiner	Author 4
Distribution χ^2 sum	300.97666	426.17741	333.56628	485.76790
LOESS residual sum	1.74409	1.51711	1.45324	1.86049

Wilks scoring represents the shape of the underlying population best for the distribution of relative scores in MEN.CL.PL. LOESS residual sums indicate that Marksteiner scores are most level across all performance bands (see table 2, figures 10 & 11).

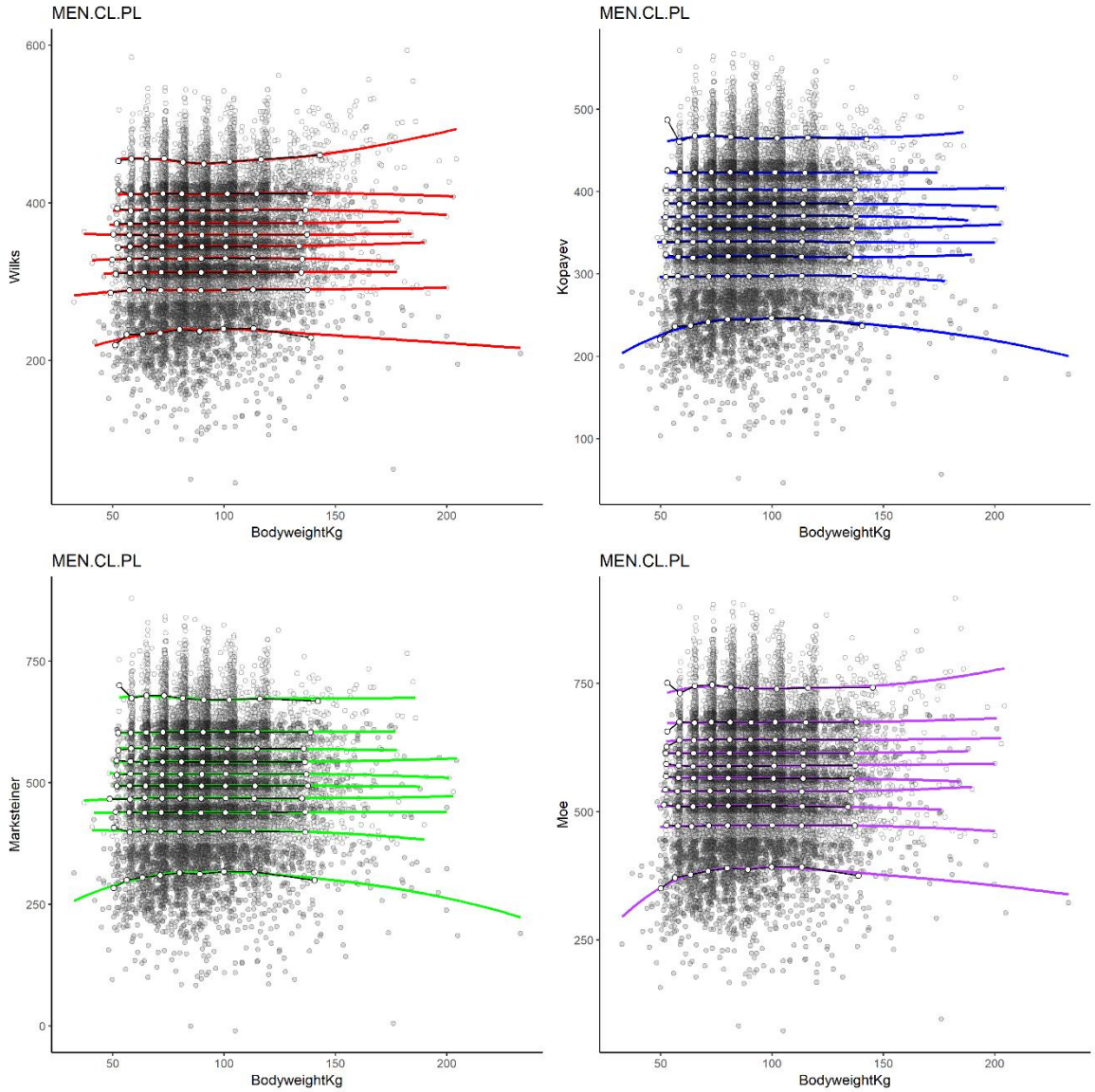

Figure 10: LOESS-plot for men's classic powerlifting

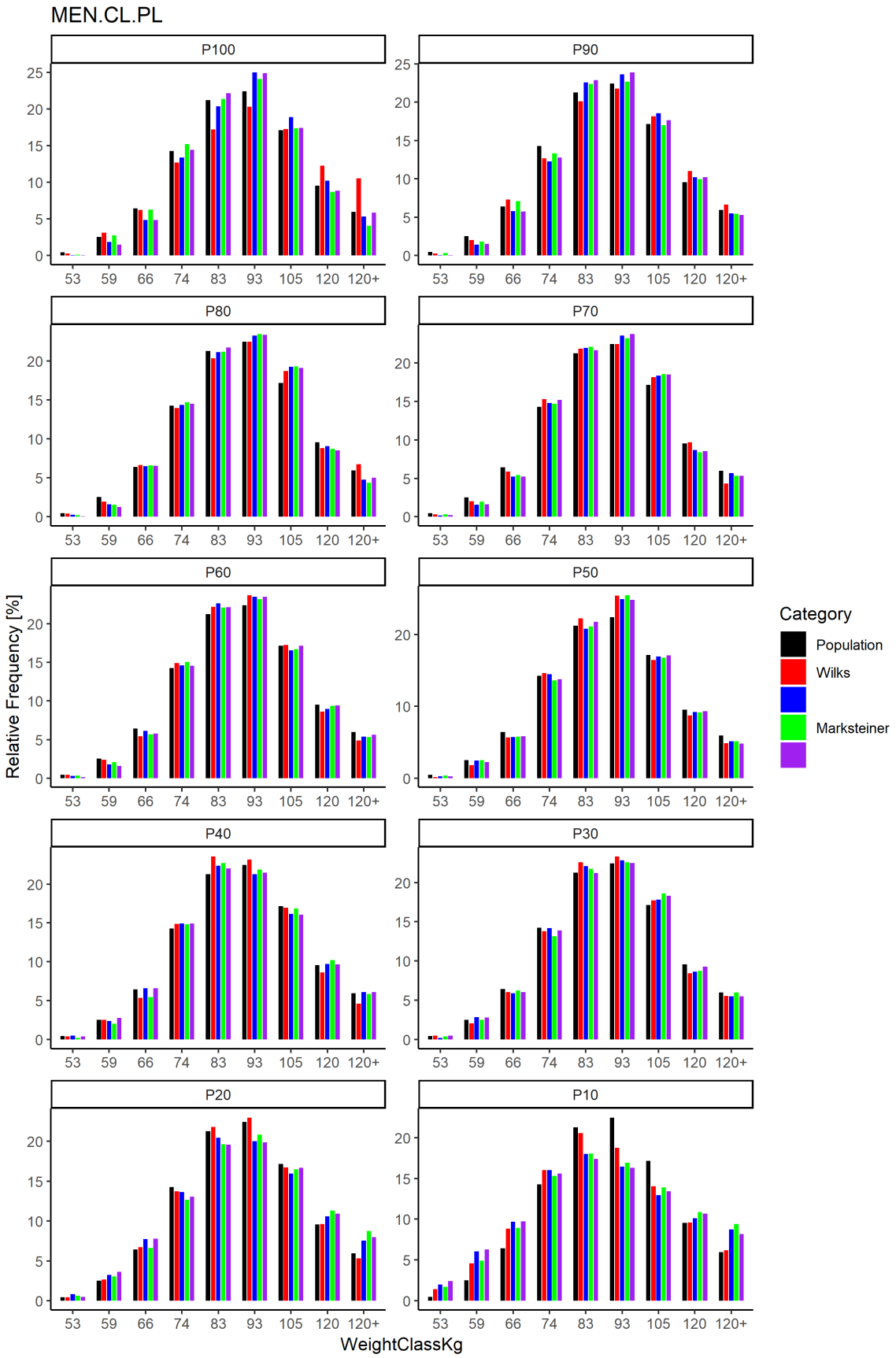

Figure 11: Distribution of relative scores across performance bands for men's classic powerlifting

5.2.3 Men's equipped bench press (MEN.EQ.BP)

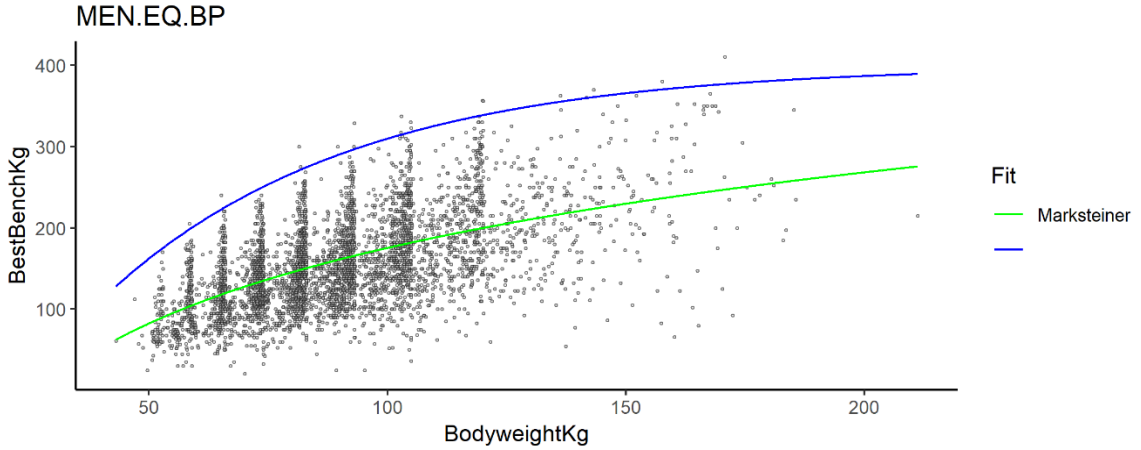

Figure 12: Model fits for men's equipped bench press

Figure 12 depicts the fits of Author 2 and Marksteiner, since Author 4 only applies to PL.

Table 3: Summary statistics for men's equipped bench press

	Wilks	Author 2	Marksteiner	Author 4
Distribution χ^2 sum	301.20240	135.44971	97.25360	317.10898
LOESS residual sum	4.58907	3.25171	2.85015	3.53149

Marksteiner scoring represents the shape of the underlying population best for the distribution of relative scores. LOESS residual sums indicate that Marksteiner scores are most level across all performance bands (see table 3, figures 13 & 14).

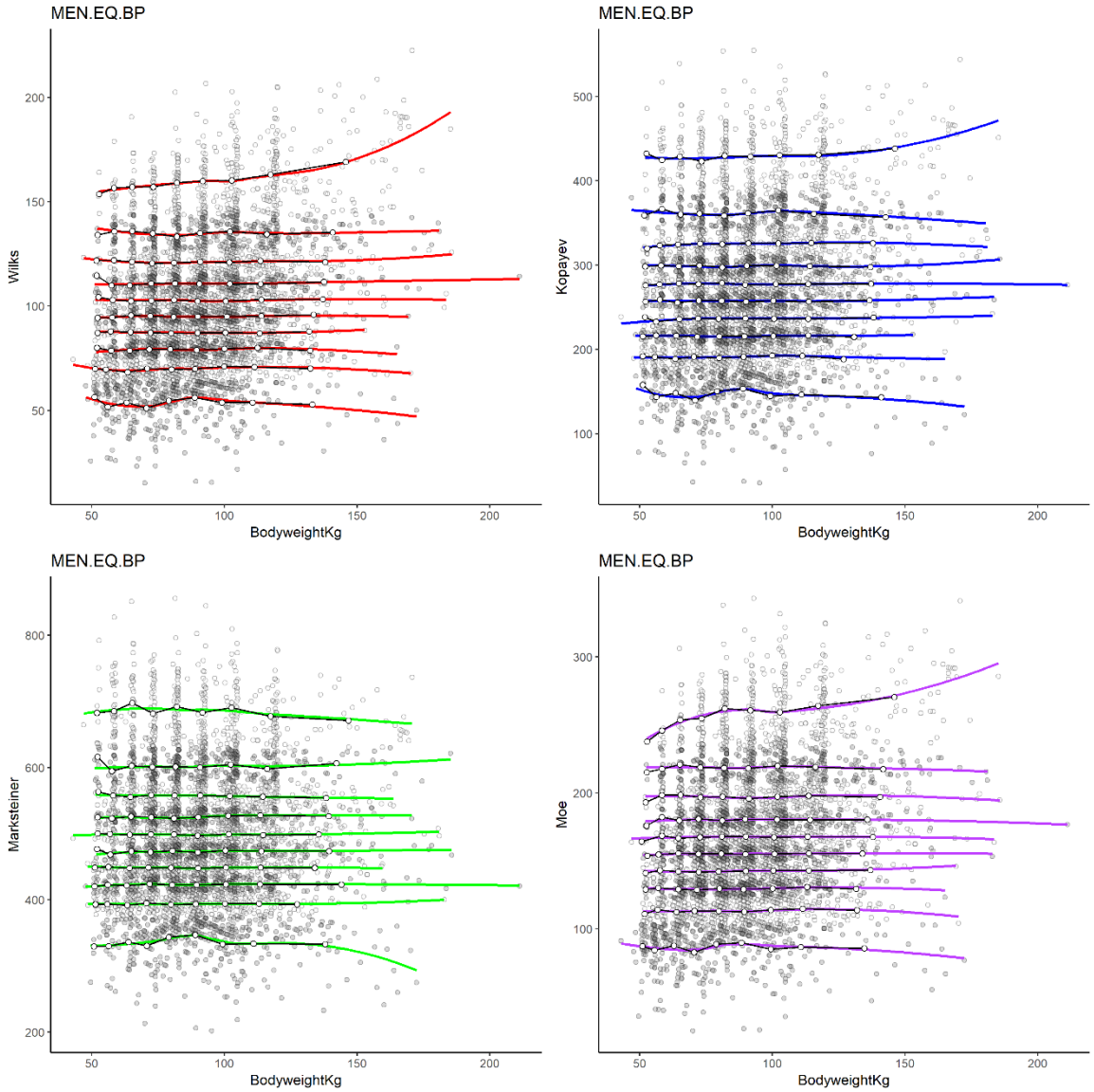

Figure 13: LOESS-plot for men's equipped bench press

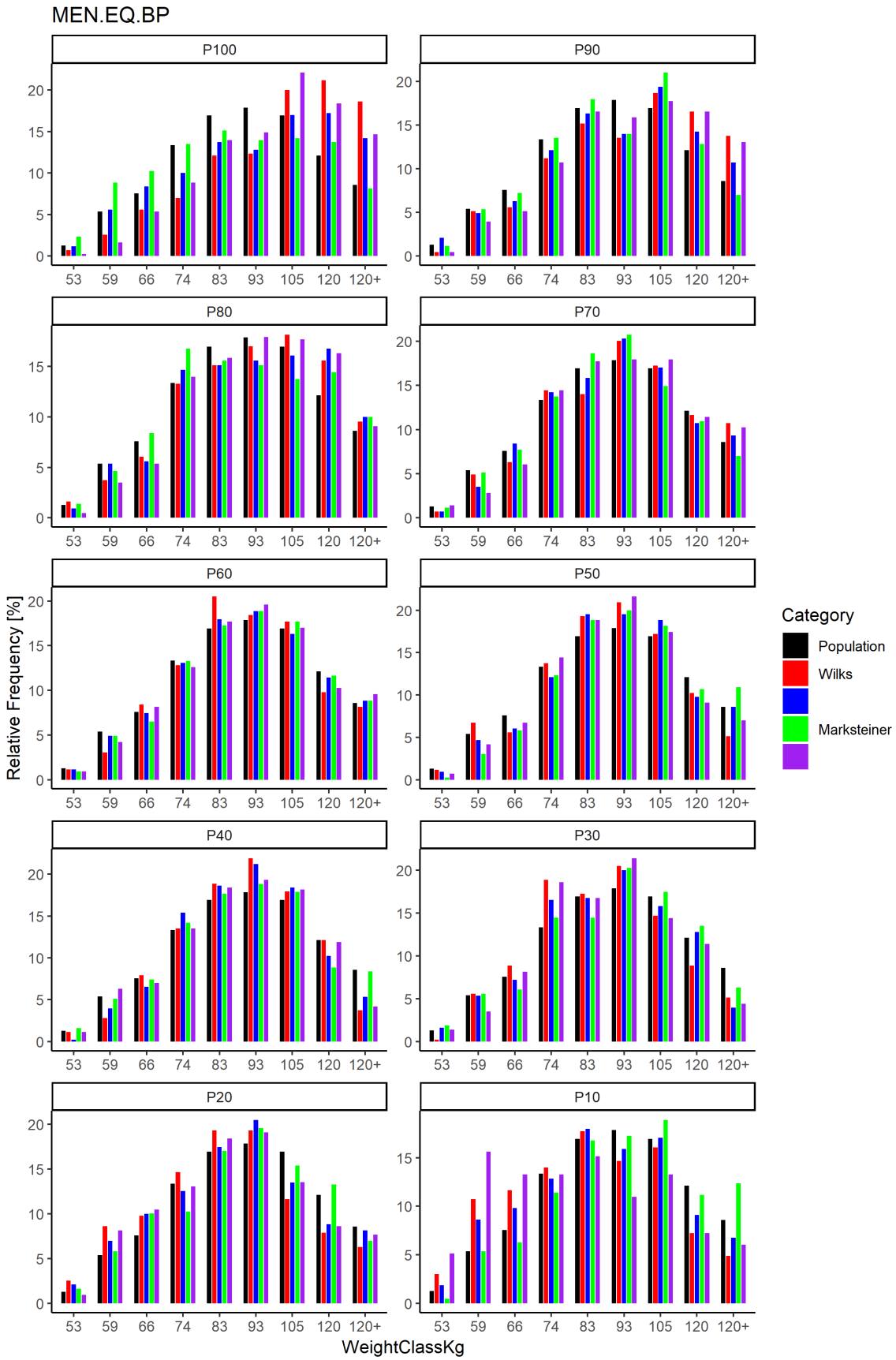

Figure 14: Distribution of relative scores across performance bands for men's equipped bench press

5.2.4 Men's equipped powerlifting (MEN.EQ.PL)

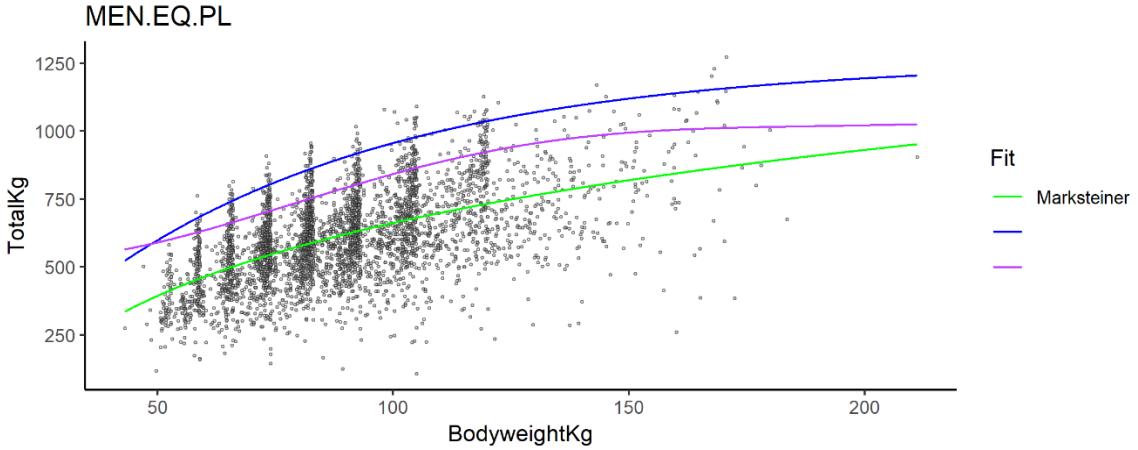

Figure 15: Model fits for men's equipped powerlifting

Table 4: Summary statistics for men's equipped powerlifting

	Wilks	Author 2	Marksteiner	Author 4
Distribution χ^2 sum	110.66918	105.93518	109.36967	148.15791
LOESS residual sum	3.12461	2.15964	2.42705	2.72819

Author 2 scoring represents the shape of the underlying population best for the distribution of relative scores. LOESS residual sums indicate that Author 2 scores are most level across all performance bands (see table 4, figures 16 & 17).

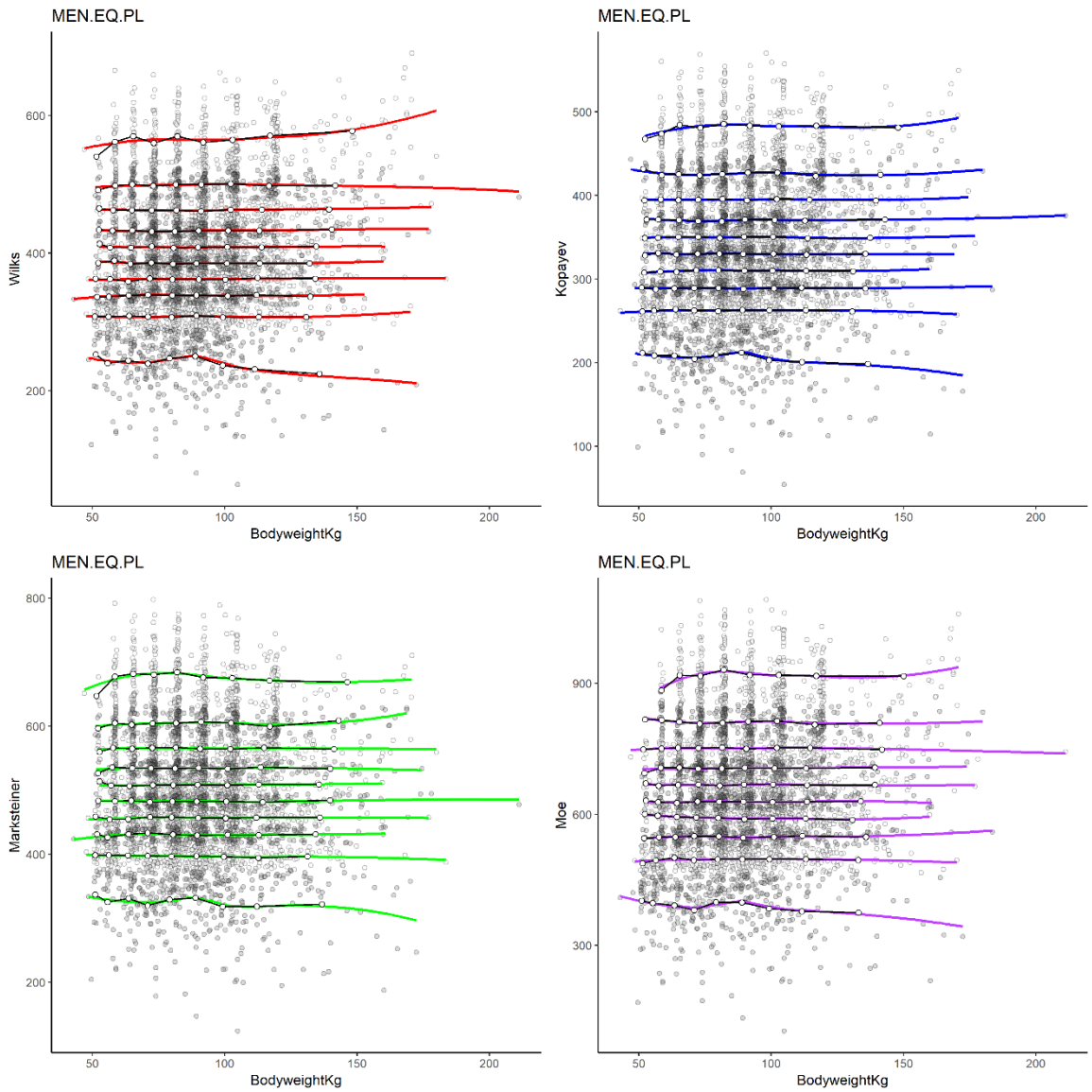

Figure 16: LOESS-plot for men's equipped powerlifting

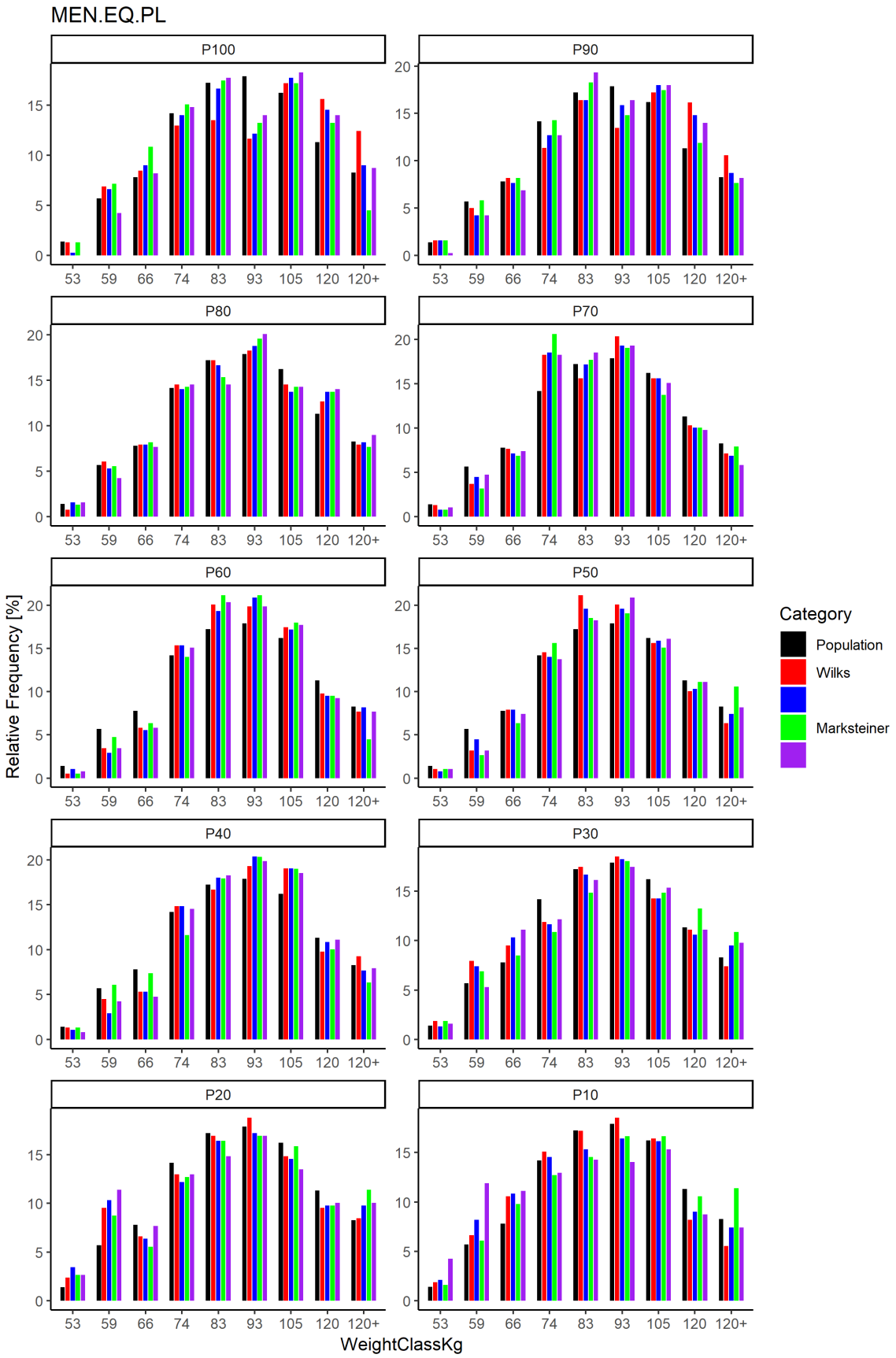

Figure 17: Distribution of relative scores across performance bands for men's equipped powerlifting

5.2.5 Women's classic bench press (WMN.CL.BP)

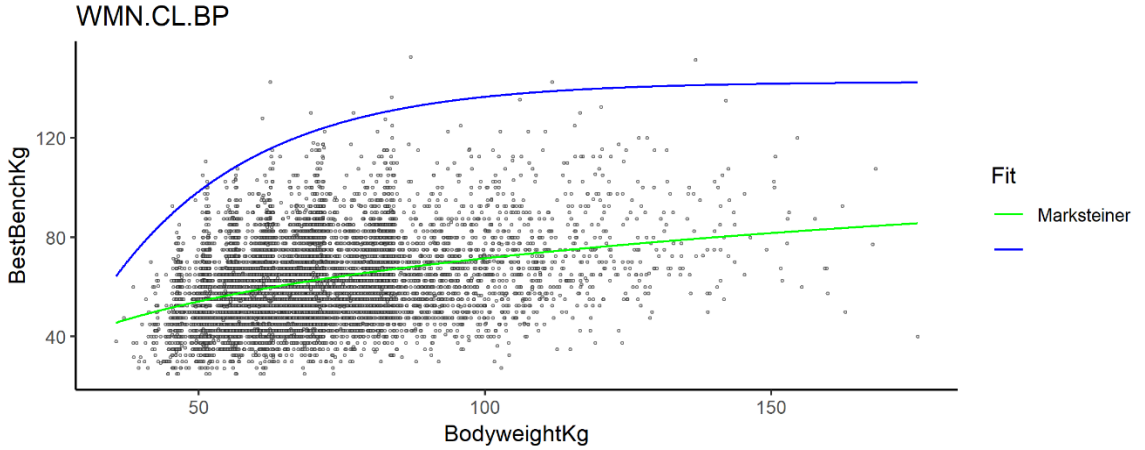

Figure 18: Model fits for women's classic bench press

Table 5: Summary statistics for women's classic bench press

	Wilks	Author 2	Marksteiner	Author 4
Distribution χ^2 sum	551.38732	190.66368	171.41021	148.97693
LOESS residual sum	2.02875	1.60337	1.73783	1.80993

For the distribution of relative scores, Author 4's scoring represents the shape of the underlying population best. LOESS residual sums indicate that Author 2 scores are most level across all performance bands (see table 5, figures 19 & 20).

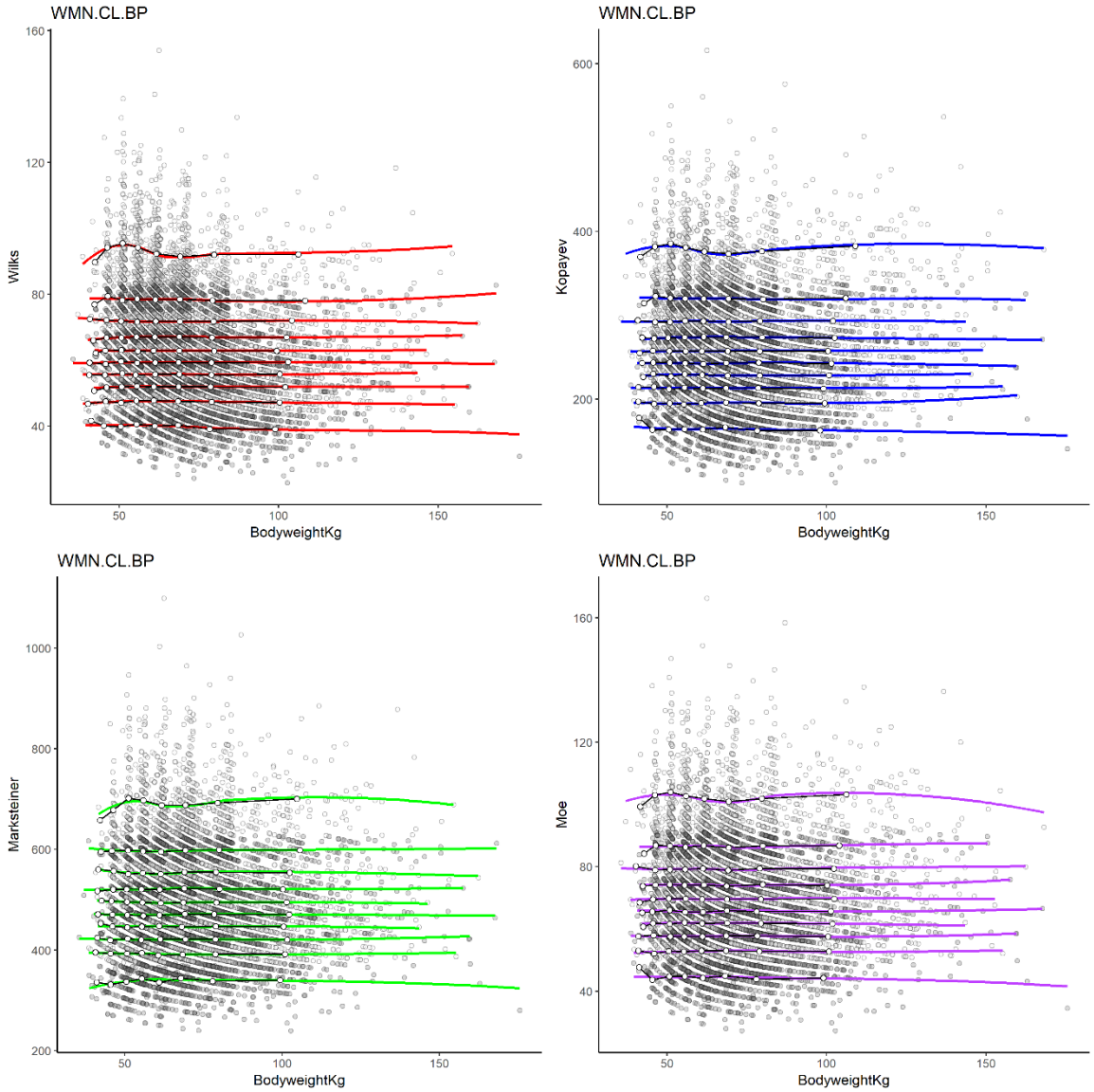

Figure 19: LOESS-plot for women's classic bench press

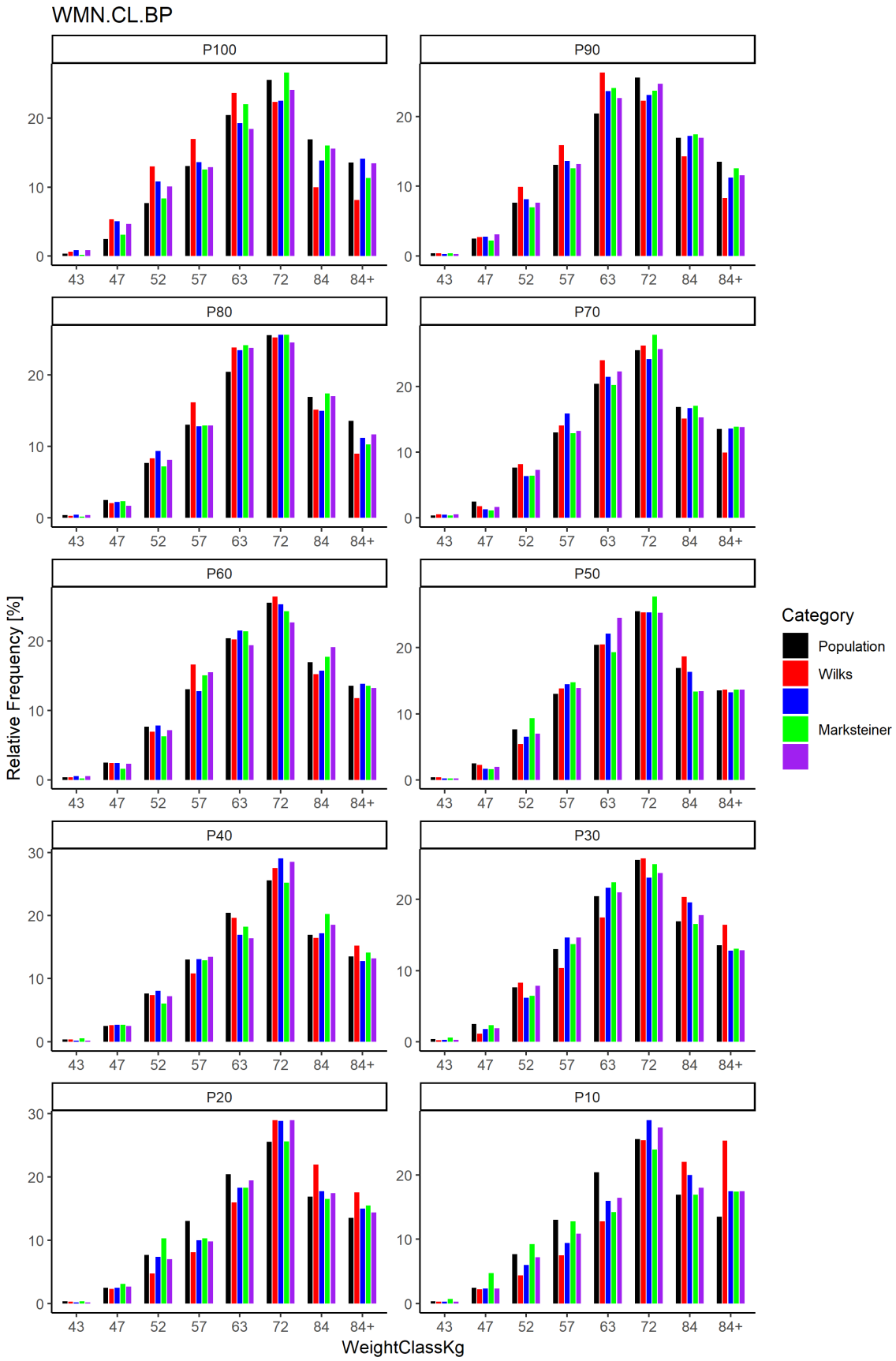

Figure 20: Distribution of relative scores across performance bands for women's classic bench press

5.2.6 Women's classic powerlifting (WMN.CL.PL)

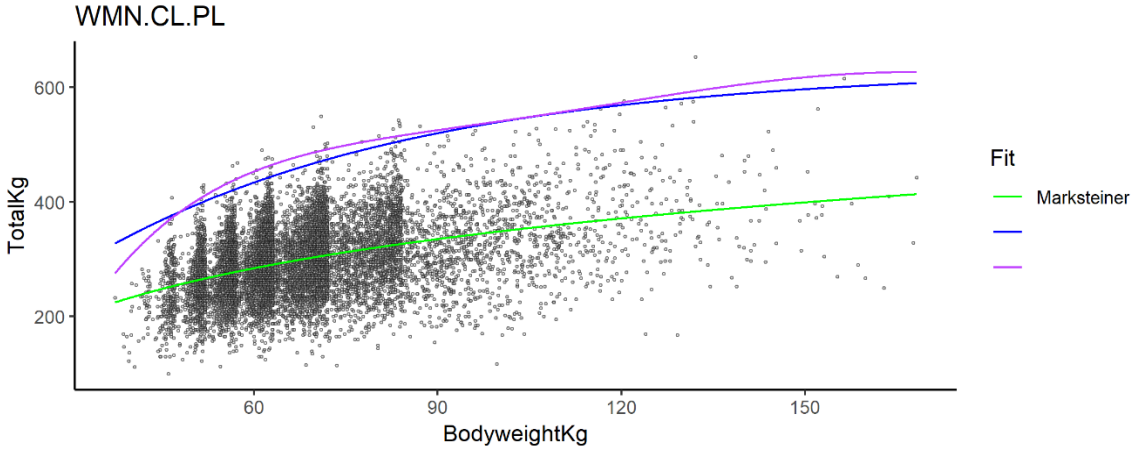

Figure 21: Model fits for women's classic powerlifting

Table 6: Summary statistics for women's classic powerlifting

	Wilks	Author 2	Marksteiner	Author 4
Distribution χ^2 sum	654.80202	158.34773	131.90024	119.80617
LOESS residual sum	2.35176	1.95914	1.97693	2.05277

Author 4's scoring represents the shape of the underlying population best for the distribution of relative scores. LOESS residual sums indicate that Author 2's scores are most level across all performance bands (see table 6, figures 22 & 23).

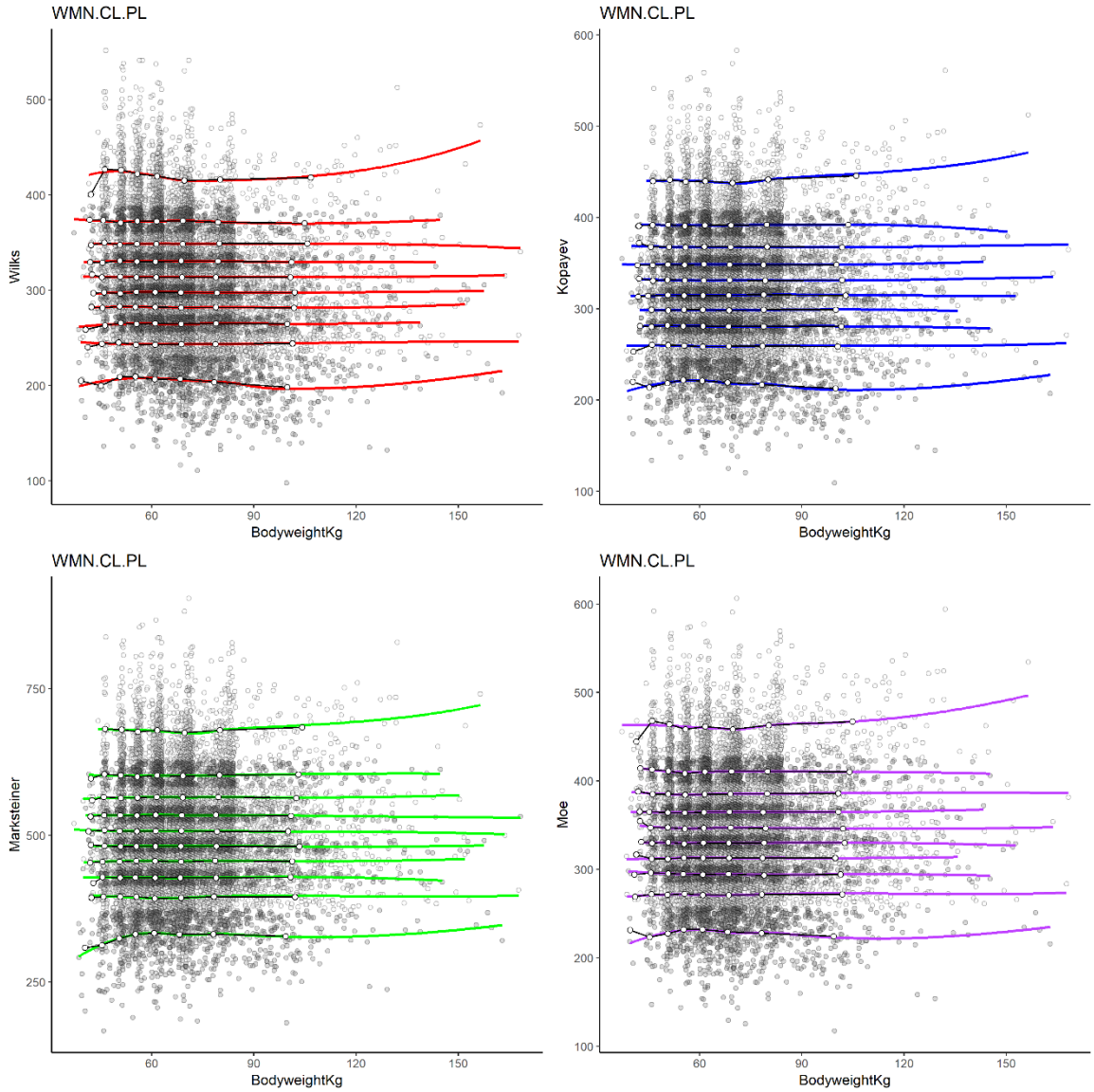

Figure 22: LOESS-plot for women's classic powerlifting

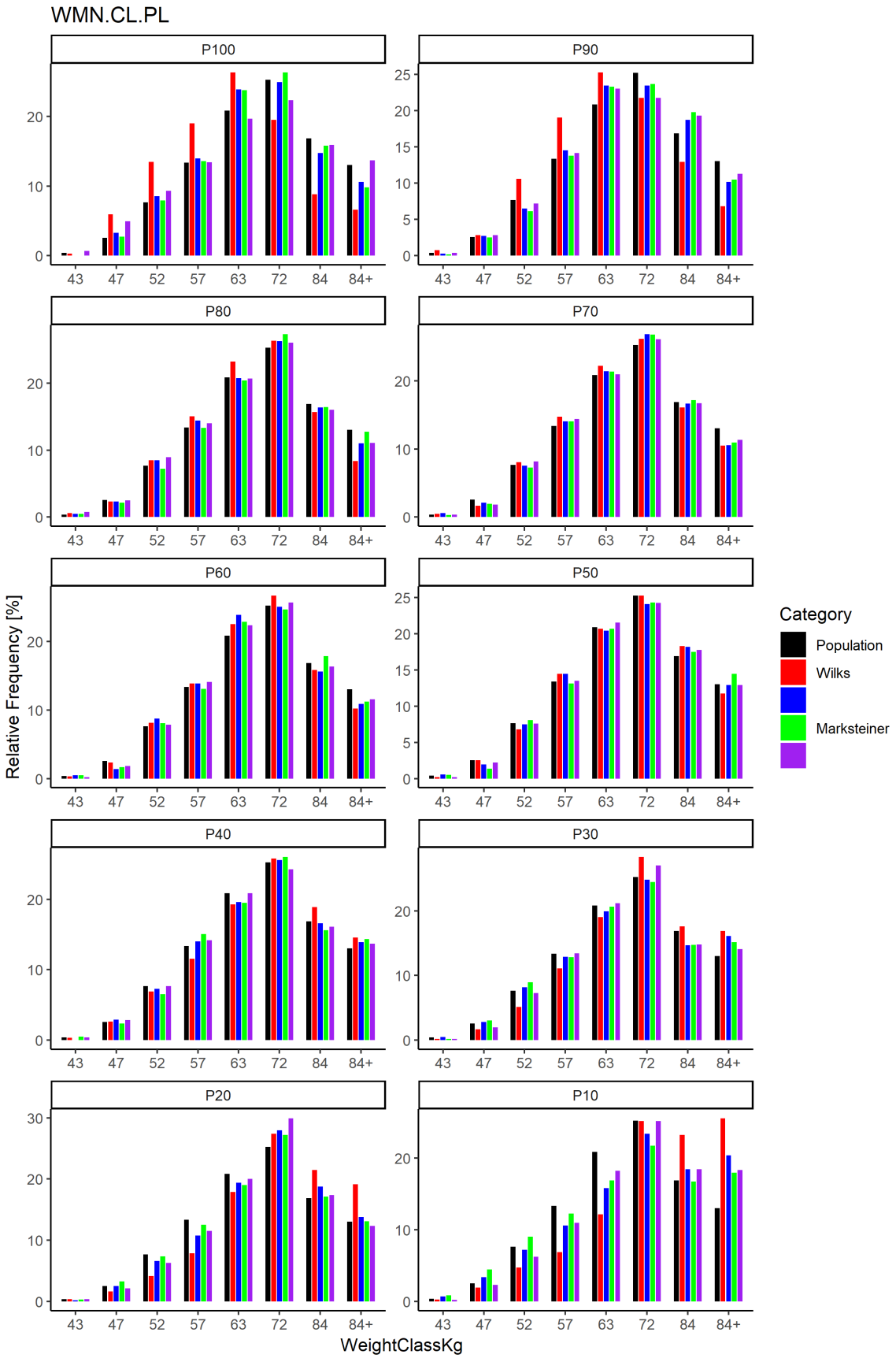

Figure 23: Distribution of relative scores across performance bands for women's classic powerlifting

5.2.7 Women's equipped bench press (WMN.EQ.BP)

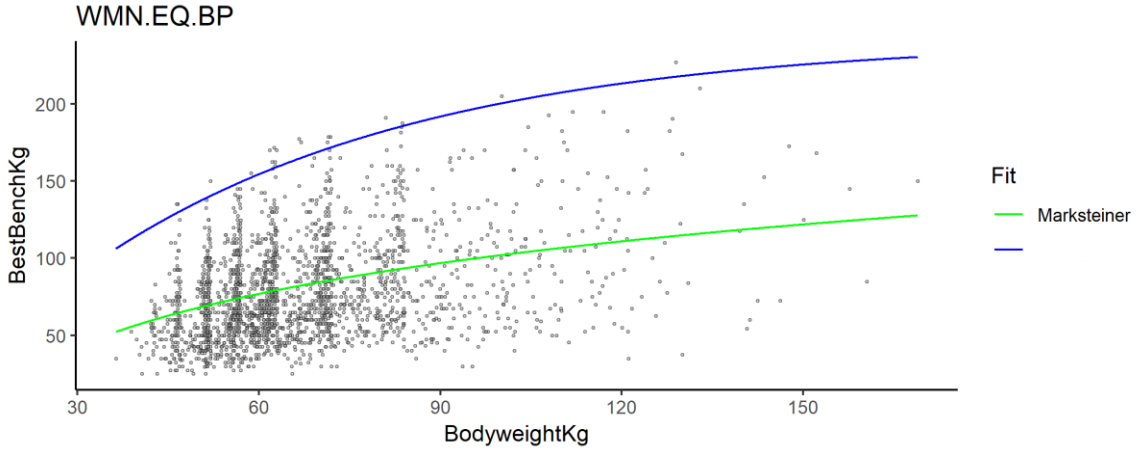

Figure 24: Model fits for women's equipped bench press

Table 7: Summary statistics for women's equipped bench press

	Wilks	Author 2	Marksteiner	Author 4
Distribution χ^2 sum	81.33261	75.43078	80.69093	64.24433
LOESS residual sum	3.33638	3.71272	4.39898	3.74472

Author 4's scoring represents the shape of the underlying population best for the distribution of relative scores. LOESS residual sums indicate that Wilks scores are most level across all performance bands (see table 7, figures 25 & 26).

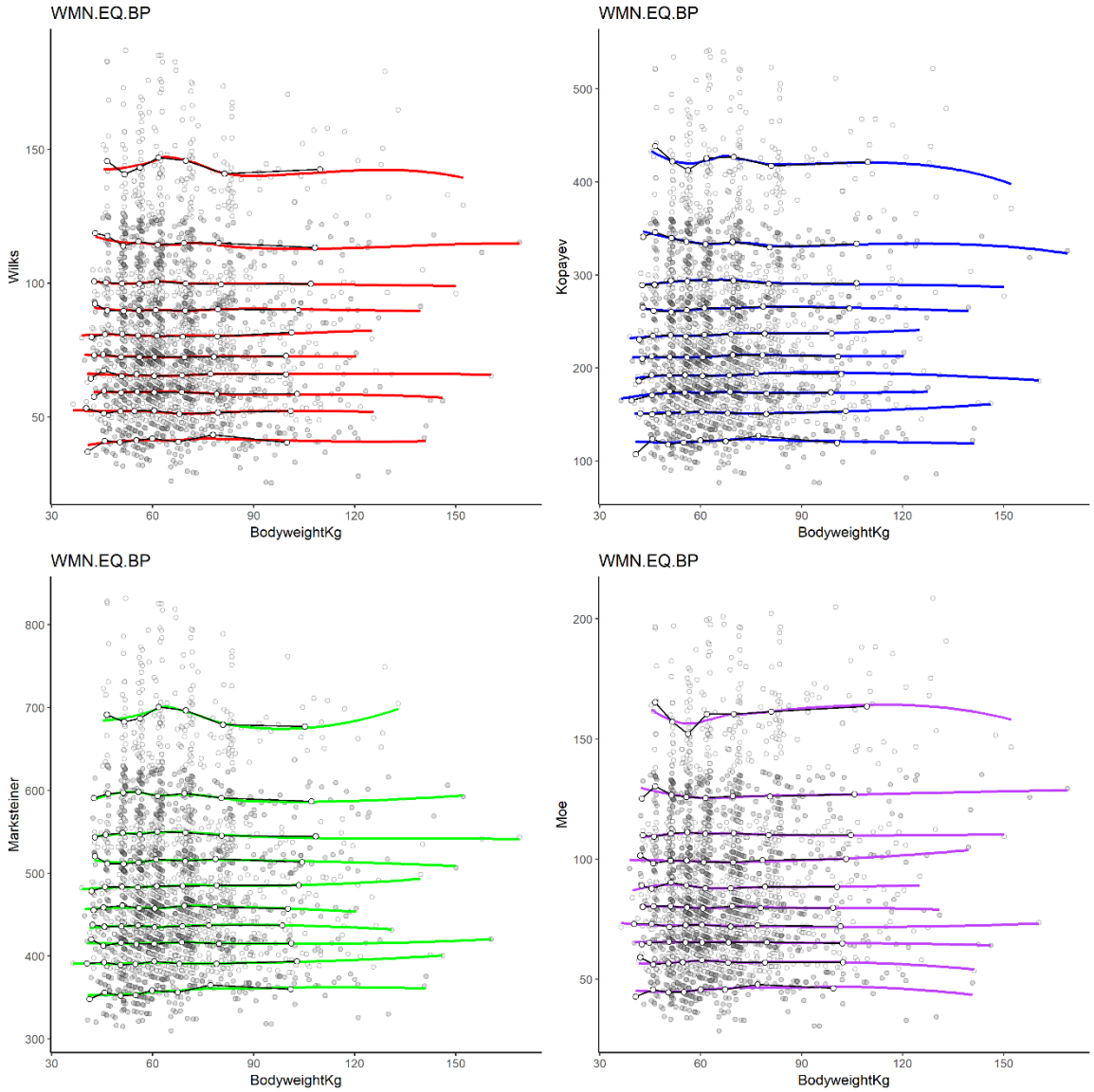

Figure 25: LOESS-plot for women's equipped bench press

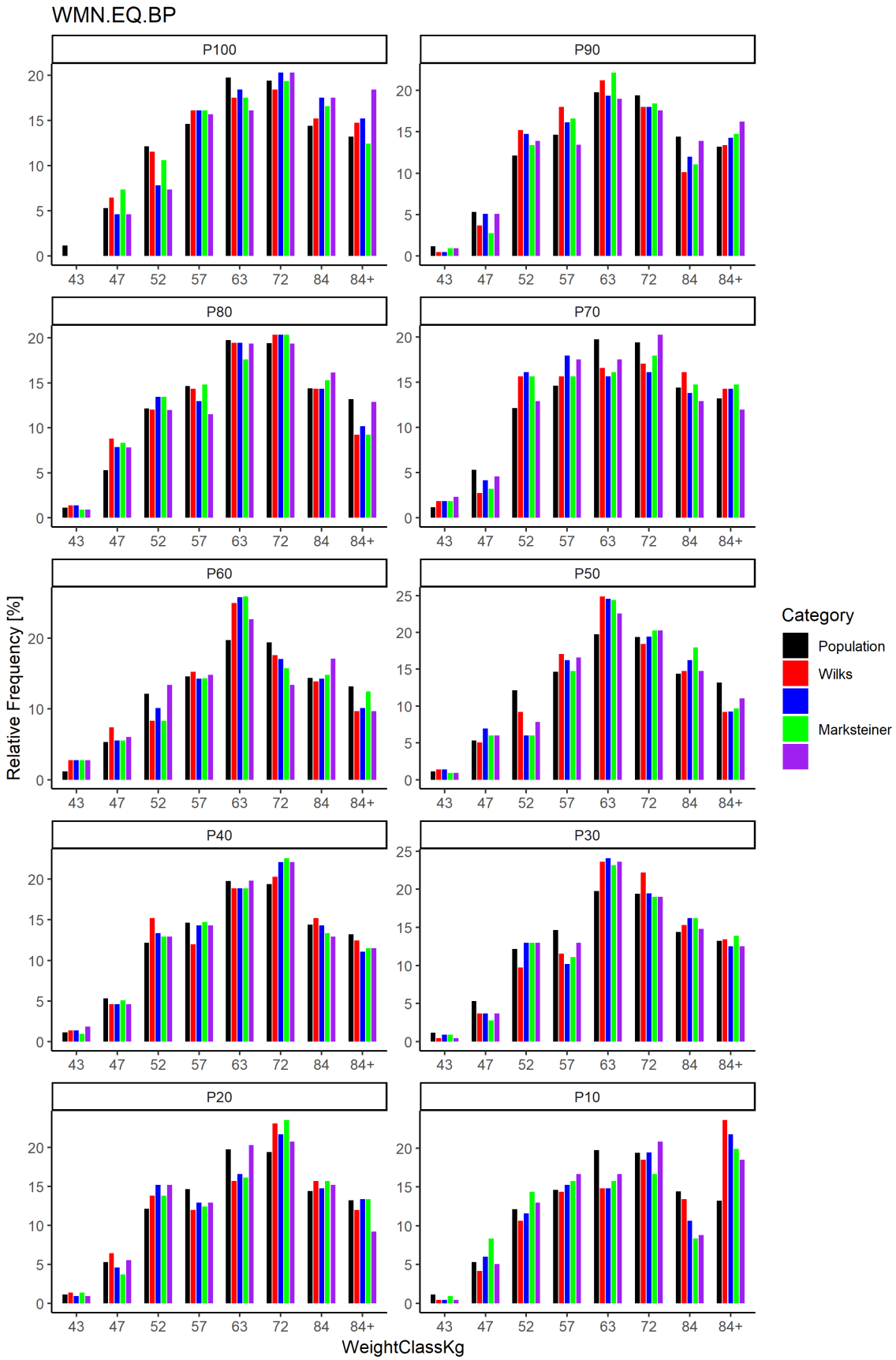

Figure 26: Distribution of relative scores across performance bands for women's equipped bench press

5.2.8 Women's equipped powerlifting (WMN.EQ.PL)

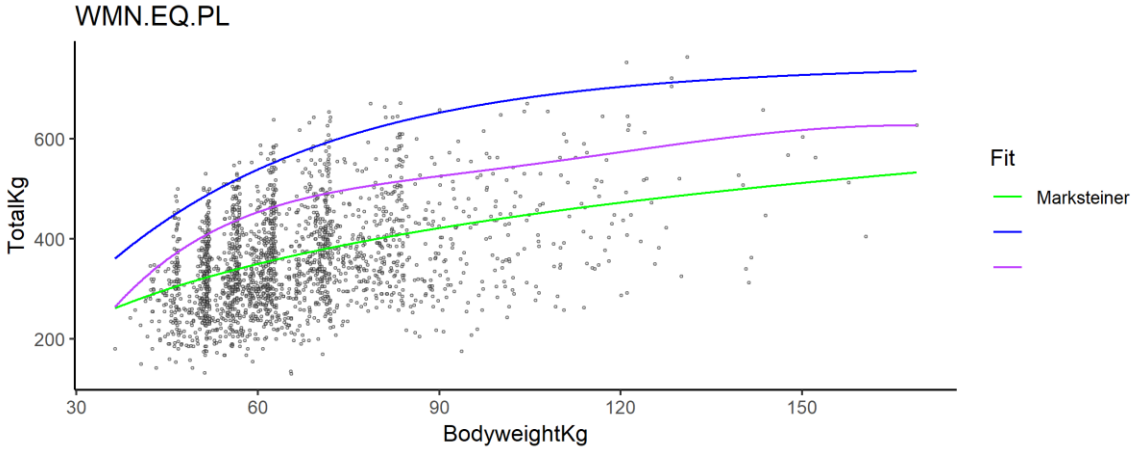

Figure 27: Model fits for women's equipped powerlifting

Table 8: Summary statistics for women's equipped powerlifting

	Wilks	Author 2	Marksteiner	Author 4
Distribution χ^2 sum	115.85353	68.72911	70.62033	75.31854
LOESS residual sum	3.68307	3.26856	3.37255	3.02682

Author 2's scoring represents the shape of the underlying population best for the distribution of relative scores. LOESS residual sums indicate that Author 4's scores are most level across all performance bands (see table 8, figures 28 & 29).

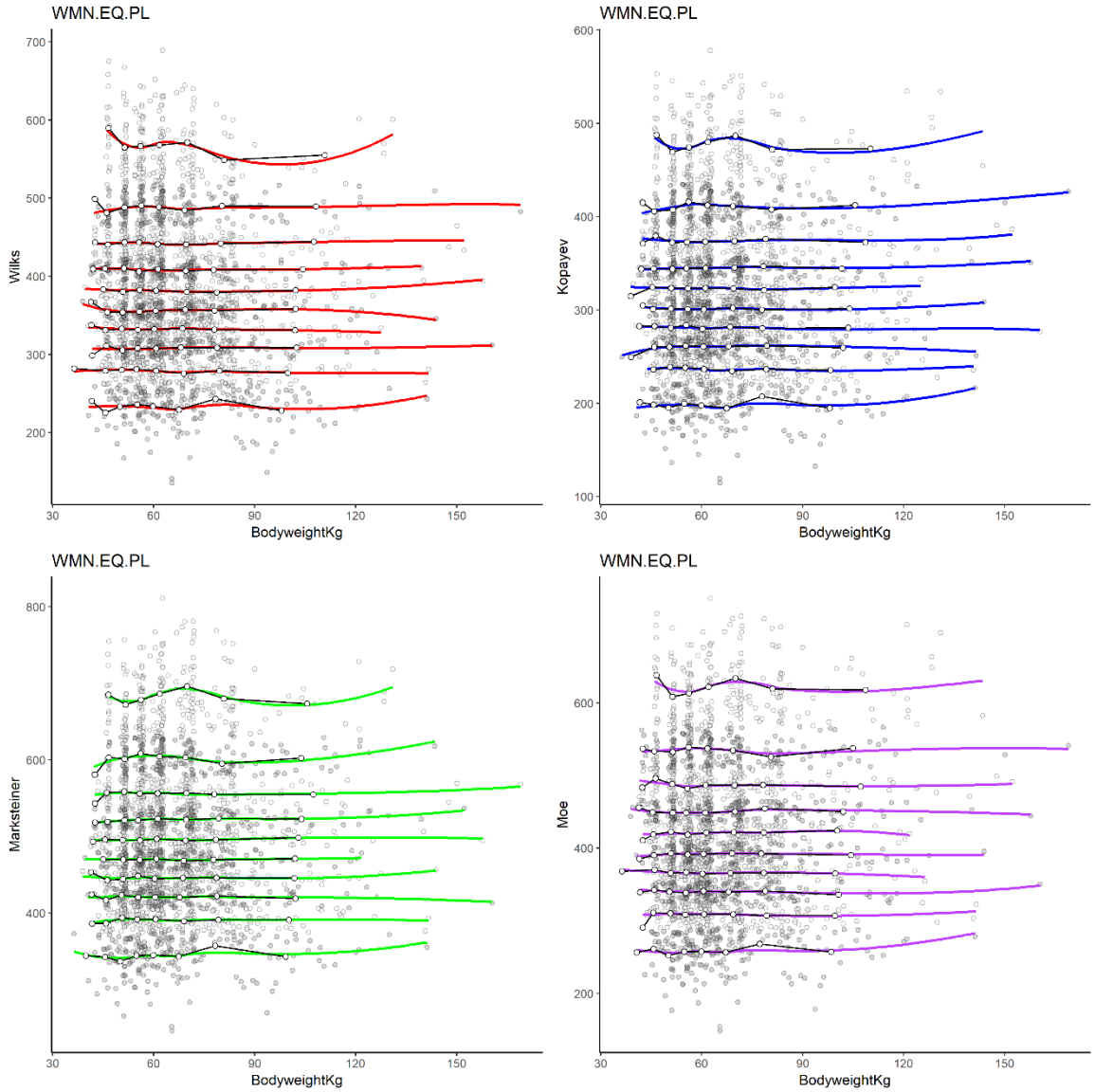

Figure 28: LOESS-plot for women's equipped powerlifting

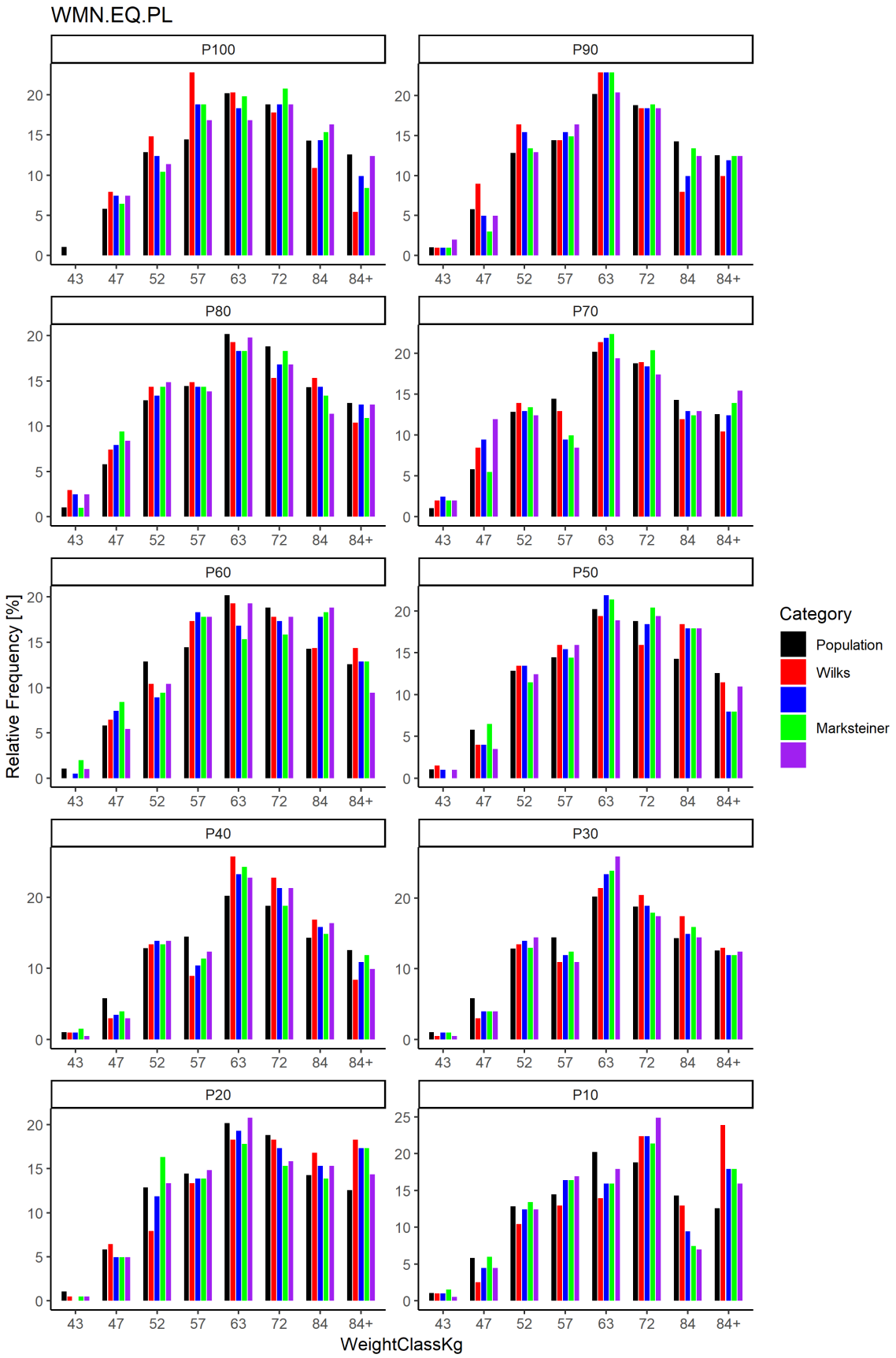

Figure 29: Distribution of relative scores across performance bands for women's equipped powerlifting

6 Conclusion

To rate the methods against each other, we will first standardize these statistics, and then calculate cumulative sums for each method.

The reviewers are not aware of any weighting preference of the IPF, so we will operate under the assumption that all subdisciplines are of equal importance. Furthermore, we will not introduce any performance band weighting within the methods. Table 9 contains the unweighted raw statistics already given in the previous sections, with their z-scores and cumulative sums in the lower area of the table. Standardization works as follows:

First, the four scores of the different methods within one domain (e. g. distribution scores for Wilks, Author 2, Marksteiner, and Author 4 for MEN.CL.BP) are scaled to a mean of zero and standard deviation of 1. These z-scores are then summed across all subdisciplines for each method. The resulting z-score sum can be found at the bottom of Table 9. The z-score sums indicate that the lowest scoring and therefore the best method is Marksteiner. Both Marksteiner and Author 2 score substantially lower (better) than Author 4 and Wilks.

Table 9: Summary of statistics across all subdisciplines

		Wilks	Author 2	Marksteiner	Author 4
MEN.CL.BP	Distr	829.96483	1014.02677	280.71897	772.38166
	LOESS	2.06503	1.91810	1.70323	1.83663
MEN.CL.PL	Distr	300.97666	426.17741	333.56628	485.76790
	LOESS	1.74409	1.51711	1.45324	1.86049
MEN.EQ.BP	Distr	301.20240	135.44971	97.25360	317.10898
	LOESS	4.58907	3.25171	2.85015	3.53149
MEN.EQ.PL	Distr	110.66918	105.93518	109.36967	148.15791
	LOESS	3.12461	2.15964	2.42705	2.72819
WMN.CL.BP	Distr	551.38732	190.66368	171.41021	148.97693
	LOESS	2.02875	1.60337	1.73783	1.80993
WMN.CL.PL	Distr	654.80202	158.34773	131.90024	119.80617
	LOESS	2.35176	1.95914	1.97693	2.05277
WMN.EQ.BP	Distr	81.33261	75.43078	80.69093	64.24433
	LOESS	3.33638	3.71272	4.39898	3.74472
WMN.EQ.PL	Distr	115.85353	68.72911	70.62033	75.31854
	LOESS	3.68307	3.26856	3.37255	3.02682
			z-scores		
		Wilks	Author 2	Marksteiner	Author 4
MEN.CL.BP	Distr	0.33752	0.92530	-1.41644	0.15363
	LOESS	1.21678	0.24660	-1.17209	-0.29129
MEN.CL.PL	Distr	-1.01064	0.46676	-0.62607	1.16995
	LOESS	0.52560	-0.66316	-0.99764	1.13520
MEN.EQ.BP	Distr	0.78558	-0.68660	-1.02585	0.92686
	LOESS	1.38988	-0.40870	-0.94875	-0.03243
MEN.EQ.PL	Distr	-0.39615	-0.63463	-0.46161	1.49239
	LOESS	1.24225	-1.08656	-0.44122	0.28553
WMN.CL.BP	Distr	1.49404	-0.39182	-0.49247	-0.60975
	LOESS	1.31475	-1.07755	-0.32134	0.08414
WMN.CL.PL	Distr	1.49711	-0.41558	-0.51747	-0.56407

	LOESS	1.46233	-0.69116	-0.59358	-0.17759
WMN.EQ.BP	Distr	0.74703	0.00077	0.66589	-1.41370
	LOESS	-1.04638	-0.19367	1.36121	-0.12116
WMN.EQ.PL	Distr	1.48840	-0.62278	-0.53805	-0.32757
	LOESS	1.26964	-0.25441	0.12796	-1.14320
Sum z-scores		**12.31775**	**-5.48717**	**-7.39753**	**0.56694**

As can be seen from most of the LOESS-Plots and bar charts, the scoring methods do not impact fairness in the middle of the performance bands. In these bands, many athletes cover a narrow range of similar performances. Relative scoring will unlikely cause significant unfairness for them, regardless of the methodology used.

The main differences between methods occur in the top and bottom performance bands. There are many possible reasons for this, for instance the philosophy of the employed fit (elite vs. average athletes) or the size and type of dataset used to set up the model(s). An entirely unweighted scoring of the statistics presumes that, although the methods perform quite differently in the top 10% of lifters, these differences will not be given special weighting in the final score. The reviewers are not aware of any reason to deviate from this equal weighting scheme.

In case the IPF follows a different philosophy, weighting the scores will change the outcome. For instance, giving more weight to the LOESS- and χ^2 scores in the top 4 performance bands (weighting factors: P100 (5), P90(4), P80(3), P70(2), P60 – P10 (1)) will change the outcome of the results given in Table 10 below.

Table 10: Weighted z-score sums across all subdisciplines, giving more emphasis to LOESS and χ^2 scores in top performing percentiles (P100 (5), P90(4), P80(3), P70(2), P60 – P10 (1))

	z-scores			
	Wilks	**Author 2**	**Marksteiner**	**Author 4**
Sum z-scores	**18.04129**	**-8.33114**	**-7.72978**	**-1.98037**

In this case, Author 2 scores lowest/best. However, if the upper two percentiles are weighted less (weighting factors: P100 (3), P90(2), P80 – P10 (1)) Marksteiners method performs slightly better again (Table 11).

Table 11: Weighted z-score sums across all subdisciplines, giving more emphasis to LOESS and χ^2 scores in top performing percentiles (P100 (3), P90(2), P80 – P10 (1))

	z-scores			
	Wilks	**Author 2**	**Marksteiner**	**Author 4**
Sum z-scores	**16.78213**	**-7.33908**	**-7.77871**	**-1.66435**

To further analyze the reason for this change in case of lower weighting we checked the performance of all methods for certain selected percentile bands. Probably most interesting is how well the top 10 % of lifters (P100) are normalized. For the top 10 % of lifters Marksteiner scores best, followed by Author 2, Author 4 and Wilks (Table 12).

Table 12: Z-score sums across all subdisciplines for the top 10 % lifters (P100)

	z-scores			
	Wilks	Author 2	Marksteiner	Author 4
Sum z-scores	**18.66197**	**-7.32236**	**-8.21277**	**-3.12683**

For the top 20 % of lifters Marksteiner again scores best (Table 13).

Table 13: Z-score sums across all subdisciplines for the top 20 % lifters (P90 & P100)

	z-scores			
	Wilks	Author 2	Marksteiner	Author 4
Sum z-scores	**18.11480**	**-7.46487**	**-7.73107**	**-2.91885**

Only for the top 30 % of lifters Author 2 performs better than Marksteiner followed by Author 4 and Wilks (Table 14).

Table 14: Z-score sums across all subdisciplines for the top 30 % lifters (P80 - P100)

	z-scores			
	Wilks	Author 2	Marksteiner	Author 4
Sum z-scores	**18.14170**	**-8.47933**	**-7.25258**	**-2.40980**

The better performance of Marksteiner in comparison to Author 2 can be mainly explained by a better consideration of the distribution of the underlying population. If the distributional component (χ^2 scores in short representing fairness) of our analysis is weighted less, e. g. half in respect to the Loess-fits (in short representing validity), Author 2 performs better for the top 10, 20 and 30 % of lifters (Table 15, Table 16 & Table 17).

Table 15: Z-score sums across all subdisciplines for the top 10 % lifters (P100) weighting χ^2 0.5

	z-scores			
	Wilks	Author 2	Marksteiner	Author 4
Sum z-scores	**13.16597**	**-6.01568**	**-4.26144**	**-2.88884**

Table 16: Z-score sums across all subdisciplines for the top 20 % lifters (P90 & P100) weighting χ^2 0.5

	z-scores			
	Wilks	Author 2	Marksteiner	Author 4
Sum z-scores	**13.38377**	**-6.12271**	**-4.78759**	**-2.47346**

Table 17: Z-score sums across all subdisciplines for the top 30 % lifters (P80 - P100) weighting χ^2 0.5

	z-scores			
	Wilks	Author 2	Marksteiner	Author 4
Sum z-scores	**13.62478**	**-6.58861**	**-4.69068**	**-2.34549**

For all lifters Marksteiner scores better than Author 2 in this weighting scheme (Table 18).

Table 18: Z-score sums across all subdisciplines for all lifters (P10 - P100) weighting χ^2 0.5

	z-scores			
	Wilks	Author 2	Marksteiner	Author 4
Sum z-scores	**9.84631**	**-4.80788**	**-5.19149**	**0.15307**

In summary, we found that both Marksteiner and Author 2 are well worked out methods, both with advantages, drawbacks, and scientific foundation. When given the choice, Marksteiner scores can be labeled as the fairer system when all subdisciplines and performance levels are taken equally into account (regardless of weighting). When focusing on elite & top 20 % lifters and weighting both components of our analysis (χ^2 & Loess scores) equally, Marksteiner again performs better than Author 2. When weighting χ^2 scores half in comparison to Loess scores, Author 2 performs better for elite, top 20, and top 30 % of lifters. However, the reviewers want to emphasize they currently see no reason to apply such a χ^2 attenuation.

Several additional aspects may change the outcome of such a scoring system. Controlling for age may be one of the factors with the biggest impact. Since age is clearly related to the strength-bodyweight relationship, age standardization could significantly alter model results and relative scoring. However, standardizing lifting performance to age has not been extensively discussed before this review. The reviewers are not aware of any ongoing discussions within the IPF that would justify age standardization at this point. Thus, it was completely omitted from this review, and datasets were left as they are, including all age groups present in the original datasets.

7 R Code used for calculating relative scores

For Marksteiner and Author 2, vector `C` contains the specific column of coefficients taken from data frame `COEF`, matching the variable `ID` (which is the filename) for each dataset with the column names. Lifter performance for PL is included in the variable `TotalKg`, for BenchPress in the variable `BestBenchKg`.

7.1 Author 2

```
COEF.AUT2 <- data.frame(MEN.EQ.PL = c(1256.96, 1440.39, 0.01566),
                        MEN.CL.PL = c(1225.49, 1108.60, 0.00940),
                        MEN.EQ.BP = c(399.49, 626.78, 0.01943),
                        MEN.CL.BP = c(344.45, 288.97, 0.00796),
                        WMN.EQ.PL = c(746.81, 1001.1, 0.02616),
                        WMN.CL.PL = c(632.84, 616.93, 0.01886),
                        WMN.EQ.BP = c(241.88, 266.63, 0.01857),
                        WMN.CL.BP = c(142.57, 321.15, 0.03967))

C <- COEF.AUT2[,which(names(COEF.AUT2)==ID)]
```

(Note that coefficient `c` for `COEF.AUT2$WMN.CL.BP` is taken from the original R output on page 39, not from table 3.4, which seems to contain an erroneous last digit.)

7.1.1 PL:

```
Author_2Points <- data$TotalKg * 500 / (C[1] - C[2] * exp(1)^(-C[3] *
data$BodyweightKg))
```

7.1.2 BP:

```
Author_2Points <- data$BestBenchKg * 500 / (C[1] - C[2] * exp(1)^(-C[3] *
data$BodyweightKg))
```

7.2 Marksteiner

```
COEF.MKS <- data.frame(MEN.CL.PL = c(310.67, 857.7850, 53.2160, 147.0835),
                       MEN.CL.BP = c(86.4745, 259.155, 17.5785, 53.122),
                       MEN.EQ.PL = c(387.265, 1121.28, 80.6324, 222.4896),
                       MEN.EQ.BP = c(133.94, 441.465, 35.3938, 113.0057),
                       WMN.CL.PL = c(125.1435, 228.03, 34.5246, 86.8301),
                       WMN.CL.BP = c(25.0485, 43.848, 6.7172, 13.952),
                       WMN.EQ.PL = c(176.58, 373.315, 48.4534, 110.0103),
                       WMN.EQ.BP = c(49.106, 124.2090, 23.199, 67.4926))

C <- COEF.MKS[,which(names(COEF.MKS)==ID)]
```

7.2.1 PL:

```
MarksteinerPoints <- 500 + (data$TotalKg - (C[1] * log(data$BodyweightKg) -
C[2]))/(C[3] * log(data$BodyweightKg) - C[4]) * 100
```

7.2.2 BP:

```
MarksteinerPoints <- 500 + (data$BestBenchKg - (C[1] * log(data$BodyweightKg) -
C[2]))/(C[3] * log(data$BodyweightKg) - C[4]) * 100
```

7.3 Author 4

7.3.1 MEN PL:

```
Author_4Points <- (842.04/(567.17331903331046-
            3.8274022516399691*data$BodyweightKg+
            0.085540790528515487*data$BodyweightKg^2+
            0.00038089371816506300*data$BodyweightKg^3-
            9.6534028292103307e-06*data$BodyweightKg^4+
            4.5466325500142581e-08*data$BodyweightKg^5-
            6.8016890038424778e-11*data$BodyweightKg^6))*data$TotalKg
```

7.3.2 MEN BP:

```
Author_4Points <- (842.04/(567.17331903331046-
            3.8274022516399691*data$BodyweightKg+
            0.085540790528515487*data$BodyweightKg^2+
            0.00038089371816506300*data$BodyweightKg^3-
            9.6534028292103307e-06*data$BodyweightKg^4+
            4.5466325500142581e-08*data$BodyweightKg^5-
            6.8016890038424778e-11*data$BodyweightKg^6))*data$BestBenchKg
```

7.3.3 WMN PL:

```
Author_4Points <- (540.34/(-817.96025420079411+
            52.980402587880548*data$BodyweightKg-
            0.85759063714003292*data$BodyweightKg^2+
            6.9435316930538046e-03*data$BodyweightKg^3-
            2.7144538222479236e-05*data$BodyweightKg^4+
            4.0708823967668707e-08*data$BodyweightKg^5))*data$TotalKg
```

7.3.4 WMN BP:

```
Author_4Points <- (540.34/(-817.96025420079411+
            52.980402587880548*data$BodyweightKg-
            0.85759063714003292*data$BodyweightKg^2+
            6.9435316930538046e-03*data$BodyweightKg^3-
            2.7144538222479236e-05*data$BodyweightKg^4+
            4.0708823967668707e-08*data$BodyweightKg^5))*data$BestBenchKg
```