Vol. VII, No. I (Volume 2022) Pages: 53– 63 ISSN (Print): 2788-5569 ISSN (Online): 2788-5445

GPSR Global Pharmaceutical Sciences Review

Cite Us

Estimating Body Fat from Abdominal Circumference and Body Mass Index in Men

Mia Md. Tofayel Gonee Manik[a]

Abstract: This secondary analysis evaluated the predictive value of adding abdominal circumference to age and body mass index (BMI) for estimating densitometry-based body fat percentage. Using 251 male records from the StatLib archive, five-fold nested cross-validation was applied across a mean benchmark, ridge regressions, random forest, and gradient boosting. Adding abdominal circumference to age and BMI reduced mean absolute error from 4.45 to 3.90 percentage points (a 0.55-point reduction; 95% bootstrap CI: 0.25–0.86). Full-panel ridge regression achieved 3.69 points MAE, while random forest and gradient boosting reached 3.88 and 3.77, respectively. The compact model overestimated the lowest body fat quartile by 4.15 points and underestimated the highest by 3.08, while adding a 2 cm measurement shift increased predicted body fat by 1.34 points. Including a single waist measurement enhances accuracy but shows sensitivity to measurement protocol and extreme-value bias.

Key Words: Body Composition, Densitometry, Nutritional Assessment, Anthropometry, Ridge Regression, Machine Learning.

Introduction

Two adults may be the same height and weight, but carry very different amounts of fat. Because body mass index (BMI) uses a numerator that includes all tissues, it maintains the balance. This simplicity is beneficial in population surveillance, but is not useful for nutritional assessment as it poses an important question that is not answered. Body mass changes may be explained as the result of fat, lean tissue, water and/or some combination of these. An anthropometric model describing body fat percentage is more precise than a body size anthropometric model and its validity depends on the accuracy of its application to individual measurements. In this work we seek to ask the question, What is the maximal accuracy that can be achieved by supplementing it with one abdominal circumference, and will the additional circumferences and flexible machine learning approach be more accurate than the compact model?

Body size and body composition are important issues to nutritional research, as the latter affects the physiological interpretation due to the former. Energy associated with lean tissue (lipid poor tissue) is different from the energy stored in the lipid rich tissue. Hall, 2008 looked at the differences related to energy deficits for weight maintenance when changing body composition, and when changing the type of body composition change. Those considerations are why people wish to know the composition of their body and they don't change the experience of a tape measure into a metabolic measurement. An estimated fat percentage in a cross-sectional study can not inform about the change in the diet or the change in a fat tissue that occurred during a nutritional intervention or whether a certain energy expenditure was required to achieve a specific result.

It is also important how the reference measurement is done, on a biochemical level. Hydrodensitometry involves the measurement of body mass and volume to estimate its density. Then it is assumed that the body is composed of fat and fat free components and a two-component equation is used to model the body with assumed densities. The

[a] Department of Biochemistry & Molecular Biology, University of Dhaka, Dhaka, Bangladesh
Email: miatofayelgonee@gmail.com *(Corresponding Author)*

URL: http://dx.doi.org/10.31703/gpsr.2022(VII-I).07 DOI: 10.31703/gpsr.2022(VII-I).07
Citation: Manik, M. M. T. G. (2022). Estimating Body Fat from Abdominal Circumference and Body Mass Index in Men. *Global Pharmaceutical Sciences Review, VII*(I), 53-63. https://doi.org/10.31703/gpsr.2022(VII-I).07

fat-free part is composed of water, proteins, glycogen and minerals, their ratio contributing to the density. A change in those ratios could result in a different relationship between tissue composition and density among individuals. Fuller et al (1992) evaluated a four component method which includes measures of body water, body minerals and minimizes the use of assumptions regarding the composition of fat free mass. Since the information analysed in this work comes from an archival endpoint which had two components in the published work, the present analysis carries over these assumptions (even though the present statistical model fits well the published data).

Anthropometry has long been used to estimate body composition because direct measurement of the relevant tissue compartments is operationally demanding. Durnin and Womersley (1974) related multiple skinfold measurements to density measured by underwater weighing and allowed prediction equations to differ by age and sex. Skinfold thickness measures local subcutaneous tissue compressed by an instrument. Abdominal circumference combines information about body frame, trunk size, multiple tissue compartments, and measurement conditions. It is reasonable to expect an abdominal measurement to add information conditional on BMI, while remaining cautious about the specific type of tissue represented by the signal. Even an accurate circumference-based estimate would leave unanswered where fat is distributed within the abdomen.

An applied data science question follows from that measurement perspective. Collecting an additional circumference entails a protocol, training, time, and potential measurement error. Those costs are easy to overlook when analyzing a historical table in which every predictor is already present. Here, the predictor sets are defined to make the cost of added information visible. The first set contains age and BMI, the second adds an abdominal measurement, and the third uses all available raw anthropometric measurements and age. The first two share the same model family so the incremental comparison changes only the available information. The third set supports a comparison of a regularized linear estimator with two nonlinear alternatives trained on the same inputs.

Data for analysis consist of the 252 records of an obtained teaching example for fitting body fat percentage to a few body measures described by Johnson (1996). That one already reflects the predictive power of circumferences. The distinction here is a matched comparison under nested cross validation, connecting the benefit of one attached measurement, with individual error, reference agreement and systematic tape offset sensitivity. Another advantage of the archival design is that a clear comparison can be made with a previous analysis based on the same measurement problem. It can't determine whether an equation will hold true for individuals recruited today and it can't conclude that an output change, made to the equation because of an intervention, indicates a change in tissue composition.

The study has three lines of inquiry. What is the decrease in held-out mean absolute error over and above age and BMI using abdominal circumference? Secondly, does a full panel or nonlinear training outperform the compact measurement set in the same source population?Secondly, are there indications that a full panel or nonlinear training does better than the compact measurement set in the same source population? Third, what is the range of stored fat %, what range of pred errors is that in, and what happens to the pred errors if an abdominal reading is systematically reported 2 cm up? The main one is a comparison of the first question. The other comparisons describe the behavior of the models; they can be seen as explorative. In addition, the archived outcome was compared to the density equation provided by the source, to analyze the consequences of large differences between the two outcomes in the table.

Materials and Methods

Source data and study design

This is a retrospective secondary study of public data provided in the StatLib body-fat data file. The archive includes data for each individual in one row and the following variables: density (underwater weighing), archived percentage body fat, age, weight, standing height, and 10 circumferences. Johnson (1996) comments that the sample consisted of 252 men and in the file ambiguous and anomalous measurements are described. In the source text, the abdominal landmark is defined as (laterally) the level of the iliac crests and (anteriorly) the level of the umbilicus. Other waist protocols were based on other areas (sites) and so the term "abdomen" is kept throughout this paper. For a secondary analysis, no recruitment was made, no measurements were performed, and no participants were contacted. Records available fall short of adding up the information required to describe the recruitment, modern representativeness and other conditions of measurement for all.

This archive includes prose and data in one text file. The rows that contained numeric data with the expected fifteen fields were extracted and the number of extracted rows was checked against the sample size that is documented. Exact conversion factors for units were used to convert weight in pounds to weight in kilograms and height in inches to height in meters. BMI was defined as mass (kg)/height2 (m2). All variables used in the analysis were supplied from the source and there was no need for imputation for the records that were parsed. A height of less than 55 inches was ignored as a "transparent screen" for the known extremely small value of 29.5 inches. The rule is important because if it is corrupted - it takes a bigger whack in the denominator of the BMI formula. It was applied before making folds and it removed only one source record. A heavier observation was retained if a transcription error could not be determined according to the recorded weight alone.

Reference phenotype and consistency audit

Archived, print body fat percentage was the primary response. The reciprocal whole-body density of two-component mixture is the weighted average of the reciprocal densities of the component masses of the mixture. The same formula results with the assumed densities of the components, namely: Percentage fat = 495/density 450 (Siri, 1956). The equation supplies an independent check on the numerical agreement of the archives two outcome related columns. It doesn't have a separate biological validation, since both quantities are derived from the same underlying measurement process. The derived quantities which were ultimate and/or contained the response or reference measurement were discarded, as well as density.

The Siri value used for the consistency audit was obtained from the density column and was set to zero (if negative) before comparison to the archived percentage. Negative values of implied fat are ignored as fat: no checking is done if they are negative. It does not treat zero typed in the archive as a discrepancy when plugging in the equation to get a slightly negative number. The one-percentage-point difference limit was selected for this reanalysis as a convenient number that is quite a bit higher than that captured by the density display to four decimal points and fat display to one decimal point. The tolerance was used to identify records for sensitivity analysis. It does not delineate any physiological levels of body fat or determine the error in measuring body fat by hydrostatic weighing.

Neither source column was unilaterally rewritten. Correcting an apparent discrepancy would require knowing whether the recorded density, the body fat percentage, or some other part of the historical record was in error. Primary estimates include disagreements after the height screen, preserving the archived endpoint. As a sensitivity analysis, the entire model fitting and evaluation procedure was rerun on the subset in which the two columns agreed within one percentage point. This reduced dataset conditions on an outcome-related property. Its error metrics therefore describe performance in a different subset and should be interpreted as a robustness check to the archived label irregularities rather than as evidence of future performance under improved laboratory procedures.

Predictor sets and estimators

The mean-only estimator assigns every test observation the training-fold mean body fat percentage. Its role is to show the error expected from using the central tendency of the source population without individual measurements. The first informative ridge model uses age and BMI. The second adds abdominal circumference to those same two inputs. The full ridge model includes age, weight, height, neck, chest, abdomen, hip, thigh, knee, ankle, extended biceps, forearm, and wrist circumferences. The full feature set represents the raw measurement panel and does not contain BMI as a derived variable. Consequently, the full-panel comparison includes a change in representation as well as in the number of measured sites.

In this context, Ridge regression is a helpful contrast, as many circumferences may have common information concerning overall size of the animal. Correlated variables can change significantly when including unpenalized separate coefficients, if they are trying to explain the same differences between the participants. Coefficient instability is reduced in a Ridge regression by penalising the sum of the squares of the coefficients only, but not the intercept. The scalings were performed within the installed pipeline, and feature means and standard deviations were only learned from the corresponding training data. The five candidate penalties were 0.1, 1, 10, 30, and 100. This design enables age, BMI, and circumference to be measured independently on the same scale and to be very influential in the model, even though measured in different units, without

having one variable shrink because it was measured in a different unit.

The random forest uses many decision trees fitted on bootstrap samples, with candidate features sampled at each split (Breiman, 2001). A tree can approximate interactions such as a change in the consequence of abdominal size at different levels of body weight without a prespecified interaction term. The forest used 180 trees and considered minimum leaf sizes of 2, 5, and 10 and maximum feature fractions of 0.5 and 1.0. The gradient-boosted tree model adds shallow regression trees in sequence to reduce discrepancies left by earlier trees, following the approach described by Friedman (2001). Its maximum tree depth was two. The grid crossed 60 and 120 estimators, learning rates of 0.03 and 0.08, and minimum leaf sizes of 4 and 10. Both nonlinear estimators used the same full input set as full ridge.

The three feature sets are pragmatic options for data collection. The age BMI equation is validated on a data set with age, height and weight only. The compact model is available with a standardized abdominal measurement. The complete group of nine additional circumference locations are provided on the full panel. Those possibilities are important when considering the difference in the process of deciding what to record when a new nutrition study is designed at the time of enrollment. It also flags an issue with the analysis. All tape measurements are taken as part of the data collecting process from the past. Interobserver and systematic differences between studies involving diverse landmarks and levels of examiner training, as well as missing tape values, cannot be observed in the experiment.

Internal validation and performance summaries

One held-out prediction per observation per model was generated by a shuffled five-fold outer split. All candidates in the Primary cohort received the same set of test blocks on the outside of the tests. Hyper parameters were selected using a mean absolute error in a shuffle four-fold search within each outer training set. Then the chosen model has been refitted on the outer training set and has been predicted on the outer test set. It is necessary to separate because the same cross validation errors used to select hyperparameters can give optimistic estimates (Varma & Simon, 2006). The (outer) split seed and the (inner) split seed were 410 and 841. Seed used by tree-estimator was 76. The consistency subset was divided then tested and analyzed again under the same conditions.

The main objective measure was mean absolute error (MAE). The units of the difference between two fat percentages are called percentage points. For instance, if the predicted value is: 24% and the archived value is: 20%, the signed error is +4 percentage points and the absolute error is 4 percentage points. More weight is given to large errors in root mean square error (RMSE). Pooled coefficient of determination is performed with reference to the average of all previously recorded results of the assessed cohort. The mean signed error represents the average directional error; positive error reflecting the overestimation. The ninetieth percentile of absolute errors refers to the error which is at the top 10% of the error distribution.

Comparisons are made with the same observations held out; each will contribute one paired difference in absolute error. A summary of the variability of those paired errors was computed by taking 10,000 bootstrapped (with replacement) observation indexes, using seed 61. Many resampled mean differences are reported (called the 2.5th and the 97.5th percentiles). These are descriptive intervals that are based on the out-of-fold predictions that have been stored. Shared models are fitted which leads to dependency among the predictions, and the bootstrap does not change the nest procedure or vary the way that the folds are allocated. Bengio and Grandvalet (2004) demonstrated that cross validation in K folds has structural limitations in terms of uncertainty estimation. The intervals specified here are constructed for the recorded error contrast and not set up to provide inferential significance from sample to sample, center to center, or for future deployments.

Errors by reference level and tape offset

Records were discriminated by the outcome in the archive and collated into four groups of equal size as possible across the range of outcomes to inspect the presence of the error of directions. For each group, the mean signed errors of scores was calculated. The boundaries of the groups are based on values observed in this sample, so they are of no diagnostic significance. Conditioning alone can lead to the shrinkage pattern if the predictor is not a perfect predictor, or even if the response is noisy. Thus, the plot is interpreted as a statement on whether a "fitted" estimate is closer to the actual extinctions in archives or not, rather than a cohort of observed

quartiles as a prospective population, or as any kind of calibration assessment. Given the value of the reference by a reader, the reader conducts an error direction analysis on the basis of the plot, according to the expected error direction.

Only the circumference of the abdomen differed (2cm larger) because of the measurement stress test in the held-out rows. Then, for each model fit, a prediction was made for both the original and changed measurements that included abdomen. The original data set was used to train the model, and the model parameters were maintained. The average within person prediction change and MAE to the same archived end-point following the perturbation are reported. The value of two-centimeter was used as an analytical stress scenario. It was not determined based on duplicate measurements or reported as a typical error of a tape protocol. The perturbation keeps the participant and reference unchanged and approximates a systematic difference in the measurements collected.

All the analysis was done using the programming languages Python, with the help from the libraries: NumPy, pandas, scikit-learn, and Matplotlib. The core estimators and validation routines are according to Pedregosa et al. (2011). No attempt was made to round / clip the regression output for model evaluation; the output was left as the natural percentage. This maintains the natural behavior of the estimators and does not give improvements in error to the post-processing. We selected the model settings to use for a small-scale benchmark, not for a comprehensive optimisation sweep. All comparisons were made to evaluate for internal purposes and no external cohort or final clinical equation were set.

Results

Cohort and audit

The archive contained 252 observations, and the primary analysis retained 251 after the height screen. Ages ranged from 22 to 81 years. Median BMI was 25.08 kg/m², with an interquartile range of 23.04 to 27.31 kg/m². The abdomen measurements had a median of 90.90 cm and a range from 69.40 to 148.10 cm. Archived fat percentage ranged from 0.0% to 47.5%, with a median of 19.2%. A zero in this archive is a value of the reconstructed reference phenotype and cannot be interpreted as absence of all bodily fat. Four records exceeded the one-point discrepancy tolerance. In their original one-based order, these were rows 48, 76, 96, and 169. Archived percentages were 5.6%, 18.5%, 17.4%, and 34.3%, respectively, whereas the recorded densities implied approximately 14.14%, 14.09%, 0.37%, and 36.25%. The sensitivity analysis retained 247 observations.

Held out prediction performance

Age and BMI lowered MAE substantially from the mean-only benchmark of 6.85 points to 4.45 points (Table 1). Adding abdomen lowered it to 3.90 points, an absolute reduction of 0.55 points and a relative reduction of about 12.3%. The conditional paired bootstrap interval for the reduction was 0.25 to 0.86 points. The same addition reduced RMSE from 5.54 to 4.81 points and increased pooled R^2 from 0.557 to 0.666. The upper portion of the absolute error distribution also improved, with a ninetieth percentile of 7.60 points for the model with abdomen versus 8.52 for age and BMI alone. Figure 1 shows the individual out-of-fold estimates and the identity line for the two compact measurement sets.

The full ridge model had the lowest primary MAE of 3.69 points. Its paired absolute-error difference relative to abdomen ridge was −0.21 points, with a conditional interval of −0.45 to 0.04 points. The random forest achieved 3.88 points, and gradient boosting achieved 3.77. Their differences relative to abdomen ridge were approximately −0.02 and −0.14 points, respectively, with conditional intervals spanning zero (Table 1). Full ridge also had the lowest RMSE and the highest pooled R^2. The rank ordering based on the point estimates alone therefore favored full ridge on the primary cohort, while the paired differences suggest limited precision for the comparisons among the anthropometric models that included abdomen.

Table 1. Primary prediction performance in 251 men

Model	MAE	RMSE	R^2	90th pct. error	Δ MAE vs abdomen	Conditional interval
Mean only	6.85	8.34	−0.004	13.25	+2.95	2.37 to 3.53
BMI ridge	4.45	5.54	0.557	8.52	+0.55	0.25 to 0.86

Model	MAE	RMSE	R^2	90th pct. error	Δ MAE vs abdomen	Conditional interval
Abdomen ridge	3.90	4.81	0.666	7.60	0.00	Reference
Full ridge	3.69	4.50	0.708	7.21	−0.21	−0.45 to 0.04
Random forest	3.88	4.71	0.680	7.83	−0.02	−0.29 to 0.24
Gradient boosting	3.77	4.62	0.692	7.61	−0.14	−0.43 to 0.15

Note. Errors are in percentage points. The interval summarizes paired absolute error differences conditional on the out-of-fold predictions.

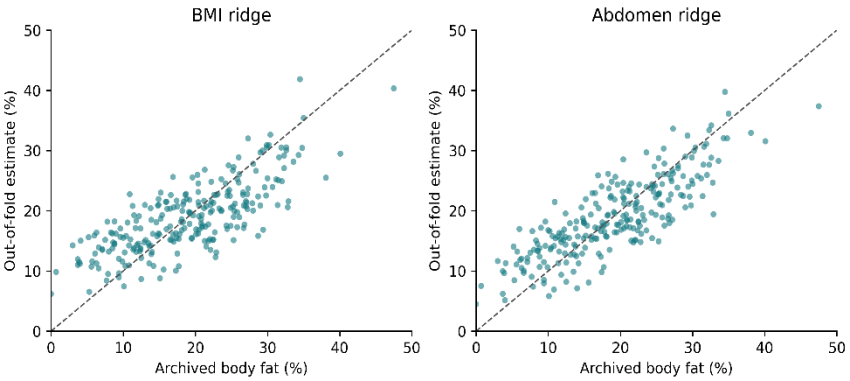


Figure 1. Archived and predicted body fat percentages. Each panel shows one held-out estimate per man. The dashed line represents exact agreement.

Reference level and robustness

Mean signed errors for the abdomen ridge model were +4.15 points in the lowest archived fat quartile and −3.08 in the highest (Figure 2). Errors in the two middle quartiles averaged +0.74 and −1.71 points. The corresponding BMI ridge values for the lowest and highest quartiles were +5.37 and −4.10 points.

Full ridge had the smallest average underestimate in the highest quartile among the fitted models shown. Every plotted model had a positive average residual in the lowest quartile and a negative average residual in the highest. These systematic differences can coexist with almost zero overall mean signed error because positive and negative deviations offset one another when averaged across all reference levels.

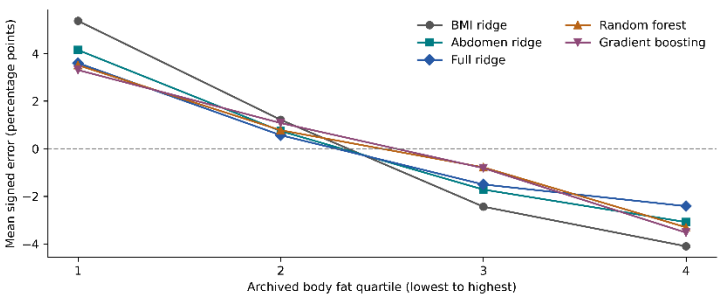


Figure 2. Mean signed errors by archived body fat quartile. Positive values indicate overestimation. Group sizes from lowest to highest were 63, 63, 63, and 62.

Excluding the four records with discordant reference fields yielded MAEs of 4.45 points for BMI ridge and 3.89 for abdomen ridge (Table 2). Gradient boosting had the lowest sensitivity-subset MAE, 3.66 points, compared with 3.71 for full ridge. Thus, the improvement associated with adding abdomen persisted under the consistency restriction, whereas the identity of the lowest-error full-panel model differed. In the tape stress scenario, a two-centimeter increase moved the mean abdomen ridge prediction upward by 1.34 percentage points and raised its MAE to 4.06. The mean shifts for the full ridge, random forest, and gradient boosting were +1.83, +1.33, and +1.23 points. Full ridge showed the largest MAE increase after the perturbation, rising from 3.69 to 4.04 points. Across all four models, the synthetic input shift produced an average prediction shift greater than one percentage point, illustrating a different aspect of performance from the baseline ordering by MAE.

Table 2. Consistency subset and abdominal measurement stress test

Model	Subset MAE (n = 247)	Mean shift after +2 cm	MAE after +2 cm
Mean only	6.87	—	—
BMI ridge	4.45	—	—
Abdomen ridge	3.89	+1.34	4.06
Full ridge	3.71	+1.83	4.04
Random forest	3.84	+1.33	4.02
Gradient boosting	3.66	+1.23	3.94

Note. All values are percentage points. The stress test used the primary cohort and applied only to models that included abdomen.

Discussion

What the abdominal measurement adds

The reduction in error when the age–BMI model acquired an abdominal measurement is scientifically interpretable without assigning a particular tissue compartment to the tape measure. BMI summarizes the amount of mass relative to the square of height. An abdominal circumference captures a separate feature of body shape. At a fixed age and BMI, people can carry their mass differently because of variation in body frame, muscularity, and the distribution of adipose tissue. A measure at the trunk can therefore supply information that body mass alone cannot identify. Ross et al. (2020) reviewed the complementary information provided by waist circumference and BMI in assessing cardiometabolic risk. The present comparison concerns agreement with an archived body fat estimate. Clinical risk and prediction of that body fat estimate are separate validation questions.

The outcome is also significant to the data collection spending plan. Here most of the attainable average error reduction was achieved after an additional tape site was added. It is reasonable to believe that a study making a choice as to how to allocate the time of the examiner is more interested in the increment of improvement achieved by the first tape measure than the decrement of part of a panel. This may or may not be a correct judgement, depending on the scientific end point. When the focus of the research question is body shape, appendicular measurements may be important regardless of their relevance for prediction of total percentage fat. When a focus of the measurement is to provide a compact small measurement for archived TBF, the small value of the wider ridge model must be balanced against the wider measurement burden and potential of protocol drift.

Both of the two tree approaches allow the relationships to vary throughout the anthropometry. This analysis did not show a clear advantage for that flexibility, when compared with the compact ridge. This can result in a largely smooth relation across the area covered by the archive, or in a small amount of data from which nonlinear interactions can be extracted, or in both. Does not assume linearity beyond the range of the data. It also does not specify which algorithm would make the best performance with larger and more diversified patient group and modern reference methods. The justifications for using a nonlinear method should be its improved performance under intended use conditions, and validation data that include the rare combination of measurements in which interactions are important.

Ridge-effects have been kept outside the scientific interpretation by purpose. Circumferences co-vary, and a model coefficient indicates the

conditional contribution of a predictor when using a penalized fitted representation. It is not able to identify a biological causal pathway. The most direct empirical evidence that this research supports for information in the abdomen is that the held-out error was altered when the abdominal measurement was added in, with the age–BMI input and the ridge framework held constant. The comprehensive panel experiment gives an answer to a different operating question, evaluating the performance of a model formed with a number of additional sites. For this reason the full ridge is based on raw height and weight data, which may mean that the performance shows differences in both information and representation. These relatively low error differences between the best models should not be given the same significance as when interpreting the larger error differences between the weaker models.

Biochemical interpretation and measurement validity

In spite of that, densitometry is a case for which a prediction target is accessible on the basis of a measurement model. The volume is estimated using underwater weighing, and volumes errors add to the density errors. There are also other values in the measurement chain, such as the residual lung volume and the density of water. Transforming from density to fat percentage is an inverse function which accentuates differences in density on the percentage scale. Within a range of densities around 1.05 g/cm³, the slope of the equation is approximately −449 percentage points per g/cm³. Thus, the estimated changes in fat for a change of 0.002 g/cm³ are about 0.9 percentage points near this density. This calculation is an example of sensitivity to the original measurement. It does not measure the size of error which occurs with underwater weighing in this population.

Another source of heterogeneity is the chemical composition of fat-free tissue. The fixed effective densities for the fat and fat-free fractions are given by a two-component equation. This measured density may mean a different % fat, if the amount of water and minerals are not at the assumed levels. Fuller et al (1992) demonstrated the use of multi-component modelling in the assessment of hydration and fat free mass density. Water balance and glycogen associated water may change in nutritional interventions and conclusions about the tissues in which the change occurred, call for a reference correct to the question. This model analyzed here only uses one measurement per person, which was archived. It thus does not provide an assessment of its sensitivity to real intrapersonal alterations nor does it provide an assessment of its ability to differentiate short-term changes in water than in lipid stores.

A useful distinction is between validity for an operational endpoint and validity for a biological quantity. Agreement with the archived fat percentage is the operational endpoint quantified in this study. A claim that the estimate reflects the amount of stored lipid in a broader population requires evidence from a suitable composition reference. A claim concerning visceral adipose tissue requires another measurement design again, because measuring total fat does not identify which anatomical depot contains it. The abdomen model makes statistical use of a feature associated with differences in shape. It cannot derive a visceral fat volume or a biochemical profile from that association, and the dataset contains no imaging or blood biomarkers with which to investigate such claims.

The audit through the Siri transformation represents a real-life lesson to be learned for applied data science. Even without causal inference, it may be possible to obtain consistency checks from related aspects of a historical data set. If there are major discrepancies, it should be noted, and, when deriving a response based on a measurement, documented separately from the source measurement prior to comparisons of algorithms. Only picking out consistent records can lead to a "cleaner" modelling exercise and also reduces rows picked. This study allows the reader to see how, by maintaining the original ordering of outcomes and providing a separate sensitivity analysis, the ordering of models is related to the records for which the provenance is most uncertain. Even if no element of internal numerical misagreement existed the assumptions of densitometry would remain.

Individual errors and use in nutrition studies

Average error can obscure the predictions most likely to be problematic for a nutritional study. For example, a proxy that compresses extremes can attenuate apparent differences between groups sampled at opposite ends of the fat distribution. The reference-level plot reveals the direction of such errors conditional on the recorded endpoint. A high archived reference value is more likely to be underestimated, whereas a low reference value is more likely to be overestimated. Some of that pattern

is expected when an imperfect predictor is compared against a noisy reference. It is therefore worth investigating how much of the compression would persist under duplicate laboratory measures and how much is attributable to regression toward the mean under squared-error fitting.

The same care should be taken when using these predictions as dependent variables. A tape-proxy estimation of body fat is used interchangeably with body weight and abdominal circumference in a trial or observation nutrition study. The inputs may move they are affected by an exposure or intervention by a path that doesn't involve a change in fat fraction. A change of position on the belly may yield a different measurement with the same amount of tissue or a change in placement of landmarks may yield a different picture, for example; the total amount of tissues remains the same but a different tape measurement is obtained. It would therefore be appropriate to consider repeatability and agreement for change scores from paired measurements at the time and interval of interest for a study primarily focused on change over time. Cancelling that exercise with crosssectional error between people cannot be done.

The comparison of the tape-offset shows how a slight adjustment in the input translates to the downstream estimate using fixed training. The modulation influence to the compact ridge model has a direct mathematical interpretation, because the same estimation change is provided by each input shift of the fold. Two participants can get different reaction from a tree ensemble when one is on the "left" side of adjacent split points while another is on the "right. Averaging such changes reflects the overall trend and sensitivity, and can obscure a spectrum of individual changes. The direction of the mean shift is significant, since if a consistent difference in the protocol between sites sees a systematic difference, even if the underlying reference distribution of the participants is similar.

Future multisite design would provide circumstances for repeated circumference measurements by the same and by separate observers. It would then be able to measure intra-observer variation, inter-observer bias and the potential effects of landmark selection. Replicate reference measurements taken for a subset of collected measurements would provide a simultaneous analysis of the action of the predictor error and the reference error. It is important to understand that holding out entire centers during development of the model is a better representation of protocol and recruitment variation for a model designed to be used throughout nutrition studies. Even if an agreement is signed for one visit, there are still additional issues regarding ‘longitudinal use’, 'recent dietary changes' and stability when measurement practices change.

Scope and limitations

No information is provided on the number of women in the historical sample or on the extent to which the findings can be generalized to other populations of different demographic make-up. The condition of body composition for prediction is developed in a particular tradition, varying according to age and sex, as well as according to the different dimensions of measurements available; further, the archive is not accessible to representative subgroup comparisons within these. The context set for the original study is not fully detailed nor are the identities of the examiners and conditions of measurement per person described. Other missing information in the data set includes dietary intake, energy expenditure, medical history, and outcomes of nutrition interventions, as well as circulating lipid measures. Thus, the nutrition relevance is based on the interpretations and measurements of body composition. The analysis is not able to establish dietary etiologies, intervention effectiveness nor biochemical relationship with laboratory markers.

The validation design is a resampling study internal to the design but with the control of the flow of information in model fitting. Records from all sources have the same similarity: tests share the same source population, test collection and test history with training folds. But, the problem with nested cross-validation is that we prevent the optimism associated with choosing hyperparameters on evaluation observations. It does not account for distribution change, from historical to current samples. Also, close comparisons can shift as the decision of resampling is made. Conditional bootstrap intervals provide information on the paired error variability inherent to this specific collection of predictions and additional cross and/or independent-sampled cohorts would be required for more extensive investigation of training-set variability.

Only a small dataset of analysis data was used and was setup on a limited modelling grid. Other types of feature engineering, flexible smooth additive models, and explicitly constrained biological models might work differently. No was required to respond

to selected contrast: Compact ridge model vs Full-panel candidates. Average absolute error was the other criterion for selecting models to fit estimators for squared-error objectives. The alternative objectives might alter tail performance or alter the central to the extreme cases. These quartiles are supposed to be representative of the response behavior in the archive and so should be treated as exploratory and interpreted with a proper mathematics of the interaction between response and residual in mind.

The findings from the study are predominantly concerns over the marginal value of obtaining an abdominal measurement for the prediction of a body composition endpoint that has been archived. The comparison also brings to light the degree of uncertainty that is present at an individual level when the overall performance is improving. For practical research use, there would need to be more evidence: the population should be predetermined, a suitable reference method should be used, the anthropometric protocol should be reproducible and valid in the population to be used. A metabolic study would also require the measurement of dietary or biochemical parameters. This study would require more measurements and analysis to separate the individual (within-person) intervention responses from stable (between-person) differences.

Conclusion

For men, the inclusion of abdominal circumference along with age and BMI decreases the held-out error in estimating the fat percentage obtained in the densitometry room many years ago. The richer ridge version provided a slightly smaller decrease, and flexible tree ensembles have not demonstrated a clear and consistent practical benefit over the more compact version. Why average error is not enough to determine fitness for use is demonstrated by directional errors at the extremes of the range of the tape measure (the reference values) and by the way it responds to a simulated shift in the tape measure. The results provide an argument to standardise the abdomen measure collected when the goal of the research is to use a low burden cross-sectional proxy for a similar end point. The result should be confirmed by at least one independent study in the same period using a different composition reference, repeated measurements from the samples, and a different population before applying it to a new population or interpreting the changes in results seen during a nutrition intervention.

Data and Code Availability

The source data and measurement notes are publicly available from StatLib at https://lib.stat.cmu.edu/datasets/bodyfat. Appendix A provides the analysis script. It reproduces the evaluation and figures using a locally saved copy of the StatLib text. The source's own data-use terms apply.

References

Bengio, Y., & Grandvalet, Y. (2004). No unbiased estimator of the variance of K-fold cross-validation. *Journal of Machine Learning Research, 5*, 1089–1105. https://jmlr.org/papers/v5/grandvalet04a.html

Breiman, L. (2001). Random forests. *Machine Learning, 45*, 5–32. https://doi.org/10.1023/A:1010933404324

Durnin, J. V. G. A., & Womersley, J. (1974). Body fat assessed from total body density and its estimation from skinfold thickness: Measurements on 481 men and women aged from 16 to 72 years. *British Journal of Nutrition, 32*(1), 77–97. https://doi.org/10.1079/BJN19740060

Friedman, J. H. (2001). Greedy function approximation: A gradient boosting machine. *The Annals of Statistics, 29*(5), 1189–1232. https://doi.org/10.1214/aos/1013203451

Fuller, N. J., Jebb, S. A., Laskey, M. A., Coward, W. A., & Elia, M. (1992). Four-component model for the assessment of body composition in humans: Comparison with alternative methods, and evaluation of the density and hydration of fat-free mass. *Clinical Science, 82*(6), 687–693. https://doi.org/10.1042/cs0820687

Hall, K. D. (2008). What is the required energy deficit per unit weight loss? *International Journal of Obesity, 32*(3), 573–576. https://doi.org/10.1038/sj.ijo.0803720

Johnson, R. W. (1996). Fitting percentage of body fat to simple body measurements. *Journal of Statistics Education, 4*(1). https://doi.org/10.1080/10691898.1996.11910505

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., & Duchesnay, É. (2011). Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research, 12*, 2825–2830. https://www.jmlr.org/papers/v12/pedregosa11a.html

Ross, R., Neeland, I. J., Yamashita, S., Shai, I., Seidell, J., Magni, P., Santos, R. D., Arsenault, B., Cuevas, A., Hu, F. B., Griffin, B. A., Zambon, A., Barter, P., Fruchart, J.-C., Eckel, R. H., Matsuzawa, Y., & Després, J.-P. (2020). Waist circumference as a vital sign in clinical practice: A Consensus Statement from the IAS and ICCR Working Group on Visceral Obesity. *Nature Reviews Endocrinology, 16*(3), 177–189. https://doi.org/10.1038/s41574-019-0310-7

Siri, W. E. (1956). Gross composition of the body. In J. H. Lawrence & C. A. Tobias (Eds.), *Advances in biological and medical physics* (Vol. 4, pp. 239–280). Academic Press.

Varma, S., & Simon, R. (2006). Bias in error estimation when using cross-validation for model selection. *BMC Bioinformatics, 7*, Article 91. https://doi.org/10.1186/1471-2105-7-91