A Machine Learning Approach to Predict Body Composition in Advanced Cancer Patients.

Pablo Cresta Morgado (pablo_crestam@hotmail.com)
A. H. Roffo Oncology Institute, University of Buenos Aires https://orcid.org/0000-0002-7502-8599
Alfredo Navigante
Bonorino Udaondo Gastroenterology Hospital
Adriana Pérez
Faculty of Exact and Natural Sciences, University of Buenos Aires

Original Article

Keywords: machine learning, body composition, advanced cancer patients, digestive cancer, predictive models

Posted Date: February 15th, 2021

DOI: https://doi.org/10.21203/rs.3.rs-200977/v1

Abstract

BACKGROUND:

Body composition and its changes affect cancer patient outcomes. Its determination requires specific and expensive devices. We designed a study to evaluate machine learning approaches to predict fat and skeletal muscle mass using daily practice clinical variables.

METHODS:

We designed a cross-sectional study in advanced gastrointestinal cancer patients. Response variables were skeletal muscle mass and body fat mass, measured by bioimpedance analysis. Predictors were laboratory and anthropometric variables. Imputation methods were applied. Six approaches were analyzed: (1) multicollinearity analysis, best subset selection (BSS) and multiple linear regression; (2) multicollinearity, BSS and generalized additive models (GAM); (3) multicollinearity, lasso to perform variable selection and GAM; (4) ridge regression; (5) lasso regression; (6) random forest. Model selection was performed evaluating the Mean Squared Error calculated by leave-one-out cross-validation.

RESULTS:

We included 101 patients under chemotherapy treatment. For skeletal muscle mass, the best approach was the combination of multicollinearity analysis followed by BSS and GAM using smoothing splines with 6 variables (albumin, Hb, height, weight, sex, lymphocytes). The adjusted R^2 was 0.895. The best approach for fat mass was multicollinearity analysis, variable selection by lasso, and GAM using smoothing splines with 3 variables (waist-hip ratio, weight, sex). The adjusted R^2 was 0.917.

CONCLUSION:

We developed the first accurate predictive models for body composition in cancer patients applying daily practice clinical variables. This study shows that machine learning is a useful tool to apply in body composition. This is a starting point to evaluate these approaches in research and clinical practice.

Introduction

Body composition is an important subject in oncology because its modifications reflect the way in which cancer affects body mass status, having implications in patients' nutrition, symptoms, and treatments. While weight loss has been considered a prognostic factor, the decrease in specific compartments could be more relevant. Losses in lean body mass (LBM) can result in a wide range of physiological impairments. This metabolically active compartment plays a role in immune function, glucose metabolism, protein synthesis and mobility [1–3]. Additionally, body composition has been linked to cancer patients' outcomes [3, 4]. Sarcopenia was associated with worse overall survival [5] and post-surgical complications [6–9]. Sarcopenia and LBM could be better to determine drug dose than body-surface area (BSA) or flat-fixed dosing [10, 11]. There is a significant association between sarcopenia and a decrease in LBM with toxicity across different oncology treatments, tumor types, and stages; suggesting an effect of sarcopenia on pharmacokinetics [3, 4, 10–16].

Several methods are available to determine body composition [17], such as anthropometry, computed tomography, and magnetic resonance [18–21], dual-energy X-ray absorptiometry (DXA) [22], and bioelectrical impedance analysis (BIA) [23]. Most of them imply high costs and are not applied in clinical practice. Anthropometry, through predictive models, has been used in some clinical fields but it has the disadvantage of having been developed only on samples of healthy people and its implementation needs training, specific supplies, and time [24]. Additionally, these predictive models have methodological issues that could be considered. For instance, all of them use linear models [25–29]. Nowadays, there is a wide set of statistical and computational tools, such as machine learning, to reach a better understanding of data (see Supplementary Material: summary of machine learning and imputation concepts) [30]. There is a variety of modern machine learning techniques which can be used to predict quantitative variables, as ridge regression, lasso regression, and generalized additive models (GAM) [30]. On the other hand, machine learning methods, which share the concept of learning from the data with machine learning, apply computational algorithms to resolve their tasks [31]. One example is random forests (RF). All these techniques, from classical to most

sophisticated ones, have a key point: the way in which they can model the data. There are more restrictive methods, as linear regression, and other flexibles ones as ridge, lasso, GAM with smoothing splines and RF [30, 32]. With regards to how to select the most important variables, some techniques were developed to solve it as best subset selection (BSS) or lasso [30]. Cross-validation, a strategy to avoid overfitting, is another important aspect to consider when a predictive model is developed. Finally, missing data is a frequent problem that could introduce bias and weaken generalizability [33]. Thus, imputation methods are useful tools to handle it [34, 35].

In conclusion, body composition has prognostic value, treatment implications, and is related to patients' symptoms and care. Specific devices, as DXA or BIA, can measure it. However, these are used in research and they are not implemented in daily clinical practice. Furthermore, neither equations nor predictive models applying clinical variables have been built in cancer patients to estimate body composition, especially considering current machine learning methods.

We performed a study to develop two predictive models to estimate body fat mass and skeletal muscle mass with clinical variables, applying several modern statistical techniques, to analyze the performance of machine learning methods and to develop a practical everyday tool.

Material And Methods

Study design and patients

Considering the impact of body composition in cancer patients, a cross-sectional study was designed to evaluate several machine learning approaches to estimate skeletal muscle mass and body fat mass using variables obtained in the clinical practice. The goal of developing a predictive model using these variables is to facilitate body composition determination in this scenario. This cross-sectional study was nested in a larger prospective study, being the data of the cross-sectional study the first measurement of the prospective one. The development of this study and the reporting process was made following the EQUATOR Network guidelines [36].

Patients aged 18 years or over were eligible for enrollment if they had histologically confirmed advanced gastric, hepatobiliary, pancreatic or colorectal adenocarcinoma, had an Eastern Cooperative Oncology Group (ECOG) performance-status score of 0 to 2, weight loss in the last six months (≥ 5%) (non-refractory cachexia) [37], had adequate hepatic, renal, and bone marrow function. Patients were ineligible if they were receiving systemic glucocorticoids, had dehydration, severe edema, or a cardiac pacemaker.

All the participants provided written informed consent. The study was approved by The Ethics Board of the Gastroenterology National Hospital "Dr. Bonorino Udaondo" Buenos Aires, Argentina, and met the recommendations stated in the Helsinki Declaration.

Variables

Dependent variables

Two variables were considered: skeletal muscle mass and body fat mass, both measured in kg. They were determined by BIA with multi-frequency (Inbody 120) according to the manufacturer specifications.

Predictor variables

Thirteen predictor variables were chosen. They were selected considering their potential association with regards to body composition. Five were anthropometric variables: height, weight. waist-hip ratio (WHr), body mass index (BMI) and BSA (by $[(\text{height (cm)} \times \text{weight (kg)}) / 3600]^{½}$). Two were sex and age.

Six were laboratory variables: hemoglobin (Hb), white-cell count (WCC), lymphocyte count (per mm^3), albumin, creatinine, and urea.

Statistical and data analysis

A descriptive analysis was developed. We evaluated the data distribution of each variable; mean or median were used according to the former, standard deviation, and first and third quartile were used, respectively.

Some key aspects need to be considered to build a predictive model or algorithm: the problem of variable selection, the collinearity between these variables, and the flexibility to model the shape of the data (see the summary on Supplementary Material). Therefore, different approaches were implemented to solve these issues. Two techniques were used to deal with variable selection: BSS and lasso regression. In this setting, multicollinearity was analyzed, considering collinear those with a Variance Inflation Factor (VIF) above 5. Associated with these variable selection methods, two regression techniques were applied. One restrictive, as linear regression, and another more flexible, as GAM with smoothing splines. On the other hand, two other different approaches were carried out separately: ridge and lasso regression. Here, a variable selection method was not added considering the capacity of each one to detect variable importance. This task is solved by adjusting variables' weight (by ridge) or dropping those less relevant (by lasso). Finally, random forests were employed, a machine learning technique which belongs to classification and regression trees. It is an accurate classification and prediction tool, and it can handle the variable importance issue.

Thus, six different approaches were applied for each response variable and their predictive performance was compared. They were: (1) multicollinearity, BSS and multiple linear regression; (2) multicollinearity, BSS and generalized additive models (GAM); (3) multicollinearity, lasso to perform variable selection and GAM; (4) ridge regression; (5) lasso regression; (6) random forest (RF). We applied smoothing splines with GAM using cross-validation to find the best degree of freedom for each variable. The tuning parameter λ for (4) and (5) was the value that produced the minimum Mean Squared Error (MSE) calculated by cross-validation. However, the value of λ for variable selection in (3) was which produces the minimum MSE plus 1 standard deviation.

Leave-one-out cross-validation (LOOCV) was used to calculate the MSE for each model to compare them. With each final model, we measured the agreement between the model and observed values by BIA according to the Bland and Altman method [38]. The 95% limits of agreement, with their 95% confidence interval, were determined. The normal distribution assumption of differences was previously checked. Besides, we plotted the difference between measurements (observed – predicted) against their mean.

Statistical significance was defined as a two-sided p-value < 0.05. Statistical analysis was performed using R software Version 3.6.3. The following packages were used: Matrix, tidyverse, carData, car, glmnet, boot, foreach, leaps, mgcv, caret, randomForest, lattice, mice.

Results

Study population and general characteristics

From August 2016 to January 2018, 101 patients with advanced upper or lower digestive adenocarcinomas were evaluated. The most frequent diagnosis was colorectal cancer (n = 52, 51.5%) and pancreatic cancer (n = 27, 26.7%). Performance Status was ECOG 0–1 in 67.3% (n = 68) and median age was 59.5. All the patients were under chemotherapy treatment with regimens based on fluoropyrimidines. Table 1 shows patients' characteristics.

Table 1
Characteristics of the patients

Variables	Values (n = 101)
Male, n (%)	67 (66.3)
Age - years, median (Q1; Q3)	59.54 (52.49; 65.48)
Performance status, n (%)	
ECOG 0	9 (8.9)
ECOG 1	59 (58.4)
ECOG 2	33 (32.7)
Tumor primary site, n (%)	
Colon and rectal	52 (51.5)
Esophageal	4 (4.0)
Gastric	18 (17.8)
Pancreatic	27 (26.7)
Body composition variables	
Height - cm, mean (SD)	165.84 (9.30)
Weight - kg, median (Q1; Q3)	71.50 (59.10; 82.50)
Skeletal Muscle – kg, mean (SD)	28.40 (6.00)
Fat mass - kg, mean (SD)	20.50 (14.05; 28.55)
Visceral fat mass - kg, mean (SD)	9.0 (6.0; 13.0)
BMI, mean (SD)	26.14 (4.90)
BSA - m^2, mean (SD)	1.81 (0.22)
Waist-hip ratio, mean (SD)	0.91 (0.07)
Laboratory variables	
Albumin - g/dl, mean (SD)	3.89 (0.23)
Creatinin - mg/dl, mean (SD)	0.83 (0.19)
Urea - mg/dl, mean (SD)	35.11 (9.33)
Hemoglobin - g/L, mean (SD)	12.77 (1.33)
White-cell count – per mm^3, mean (SD)	7039 (1872)
Lymphocytes count – per mm^3, median (Q1; Q3)	1620 (1095; 1996)

References. SD: standard deviation, Q1: first quartile; Q3: third quartile; BMI: body mass index.

There were 1.8% missing data (Figure S1). Lymphocyte has the highest percentage of missing values (14.9%), but the rest below 3%. The predictive mean matching method with 5 iterations was applied to impute each variable considering the same dataset: both response variables and predictors [35, 39].

Skeletal muscle mass predictive model

Considering the six approaches described previously, test MSE over the validation set was presented in Fig. 1A and in figure S2 to S6 of Supplementary Material. Table 2 shows the role of each variable for every approach. The best model was obtained with approach 2 according to what will be described below.

Table 2
Approaches for skeletal muscle mass

Approach [a]	Height	Weight	WHr	BMI	BSA	Sex	Age	Hb	WCC	LC	Alb	Cr	Urea
Appr1 (8), coeff	**0.25**	**0.19**	-3.86			**3.76**	-0.02	**-0.43**		**0.0006**	**3.00**		
Appr1 (7), coeff	**0.26**	**0.18**				**3.81**	-0.02	**-0.44**		**0.0007**	**3.00**		
Appr1 (6), coeff	**0.27**	**0.17**				**3.63**		**-0.39**		**0.0006**	**3.12**		
Appr1 (5), coeff	**0.26**	**0.17**				**3.59**		**-0.38**			**3.28**		
Appr2 (8), edf	**0.98**	**1.00**	2.21			**3.98**[b]	2.9e-9	**6.69**		**1.28**	**0.88**		
Appr2 (7), edf	**1**	**1**				**4.23**[b]	1	**6.81**		**1**	**1**		
Appr2 (6), edf	**1**	**1**				**4.09**[b]		**6.68**		**1**	**1**		
Appr2 (5), edf	**1**	**1**				**4.05**[b]		**6.23**			**1**		
Appr3 (4), edf	**2.20**	**0.97**				**3.40**[b]					**0.87**		
Appr4 (13), coeff	0.21	0.08	-2.51	0.02	7.15	3.30	-0.02	-0.30	1.8e-5	5.4e-4	2.72	1.16	-0.004
Appr5 (13), coeff	0.19	0.00	0.00	0.00	13.1	3.12	0.00	-0.09	0.00	3.2e-4	2.22	0.00	0.00
Appr6 (13), %IncMSE	24.9	15.3	5.86	7.15	21.1	15.5	3.19	1.20	1.07	1.78	5.13	2.30	-0.47

References.

[a] The six approaches (Appr1-6) with different combinations of variables are displayed. The number of variables included is indicated between brackets. Coefficients (coeff) are shown for Appr1, 4, and 5; effective degrees of freedom (edf) are displayed for Appr2. In Appr1 to 3, significant coefficients ($p < 0.05$) are identified in bold. Appr6, corresponding to random forests, is accompanied with the percentage of increase for MSE.

[b] According to Appr3, for sex variable the coefficient, instead of edf, is presented.

WHr: waist-hip ratio; BMI: body mass index; BSA: Hb: hemoglobin; WCC: white-cell count; LC: lymphocyte count; Alb: albumin; Cr: creatinine.

High VIF values were observed for BSA (586.99), BMI (116.99), height (72.68), and weight (537.95). After excluding BSA and BMI, VIF values were below 5 for the remaining 11 variables. In the next step, BSS was applied with these 11 variables.

BSS showed four sets of variables with similar values for R^2 (figure S2). These sets were: with 5 variables (albumin, Hb, height, weight, sex) ($R^2 = 0.8771$), with 6 (same plus lymphocytes) ($R^2 = 0.8814$), with 7 (same plus WHr) ($R^2 = 0.8824$) and with 8 (same plus age) ($R^2 = 0.8818$). Although the highest adjusted R^2 was obtained for the combination of 7 variables, since the four combinations had close values, all of them were analyzed. Therefore, the lowest test MSE was obtained applying GAM using smoothing splines with the combination of 6 variables. In this setting, the best model found was the result of the combination of

linear models (1 effective degree of freedom (edf)) for albumin ($p = 0.001$), height ($p < 0.0001$), weight ($p < 0.0001$), and lymphocytes ($p = 0.019$); a positive parametric coefficient for sex (male) ($p < 0.0001$), and a smoothing splines model with 6.68 edf for Hb ($p = 0.013$). The adjusted R^2 obtained for this model was 0.895.

In Fig. 2A, predictive values and observed values for skeletal muscle were plotted. Dots are close to the line and homogeneously distributed to both sides of it. The plot of differences against the mean (Fig. 2B) shows a random distribution of each observation and no relationship was seen between discrepancies (difference) and the values (mean). We formally examined this relationship with the Spearman's rank correlation coefficient, which is 0.132, proving no correlation ($p = 0.19$). The 95% limits of agreement were − 4.20 and 4.12 (Fig. 2B)

Body fat mass predictive model

Out of the six analysis approaches the lowest MSE, calculated on the validation set, was obtained for approach 2 with multicollinearity analysis, BSS, and GAM with 5 variables (MSE 11.01) (Fig. 2A and figure S7 to S11 in Supplementary Material). However, an MSE value of 11.04 was the result of approach 3: multicollinearity analysis, variable selection by lasso regression, and GAM with 3 variables (figure S10). This model was selected as the best regression model, considering the parsimony principle and the small difference between both MSE values. This approach is described below. In Table 3, similar to skeletal muscle, all the approaches and their variables are displayed.

Table 3
Approaches for body fat mass

Approach [a]	Height	Weight	WHr	BMI	BSA	Sex	Age	Hb	WCC	LC	Alb	Cr	Urea
Appr1 (8), coeff		**0.30**	**89.5**			**-7.36**	0.06	**0.61**	-0.0003	-0.0008	-2.67		
Appr1 (7), coeff		**0.30**	**88.9**			**-7.37**	**0.07**	0.53		**-0.001**	-2.41		
Appr1 (6), coeff		**0.30**	**89.5**			**-7.54**	**0.07**	0.48		**-0.001**			
Appr1 (5), coeff		**0.30**	**91.2**			**-7.02**	**0.06**			**-0.001**			
Appr2 (8), edf		**4.65**	**2.80**				0.09	0.16	1.05	0.50	0.03		
Appr2 (7), edf		**4.66**	**2.71**				0.60	2.1e-8		**1.16**	9.9e-8		
Appr2 (6), edf		**4.66**	**2.71**				0.60	1.8e-8		**1.16**			
Appr2 (5), edf		**4.78**	**2.77**				1			**1.23**			
Appr3 (3), edf		**4.47**	**2.84**			**-5.88**[b]							
Appr4 (13), coeff	-0.002	0.12	71.7	0.50	5.28	-6.02	0.049	0.53	-0.0002	-0.001	-2.55	-0.15	0.01
Appr5 (13), coeff	0.00	0.20	83.9	0.33	0.00	-5.68	0.03	0.29	-0.0002	-0.0008	-1.68	0.00	0.00
Appr6 (13 v), %IncMSE	6.88	12.5	28.6	20.3	9.83	7.37	-0.68	0.64	1.95	0.54	2.35	1.34	-0.28

References.

[a] The six approaches (Appr1-6) with different combinations of variables are displayed. The number of variables included is indicated between brackets. Coefficients (coeff) are shown for Appr1, 4, and 5; effective degrees of freedom (edf) are displayed for Appr2. In Appr1 to 3, significant coefficients ($p < 0.05$) are identified in bold. Appr6, corresponding to random forests, is accompanied with the percentage of increase for MSE.

[b] According to Appr3, for sex variable the coefficient, instead of edf, is presented.

WHr: waist-hip ratio; BMI: body mass index; BSA: Hb: hemoglobin; WCC: white-cell count; LC: lymphocyte count; Alb: albumin; Cr: creatinine.

According to the previous analysis, BSA and BMI were removed by multicollinearity, and the remaining 11 variables were used by lasso for variable selection. Three had its coefficient value different to zero: WHr, weight, and sex. The result of GAM analysis was a model including a negative parametric coefficient for sex (male) ($p < 0.0001$) and smooth terms for WHr (with 2.84 edf, $p < 0.0001$) and for weight (with 4.466 edf, $p < 0.0001$). The adjusted R^2 obtained for this model was 0.917.

Finally, Fig. 2C shows the scatter plot of the observed values and predictive values of body fat mass. Dots are narrowly disposed around the equality line. The plot of differences against the mean (Fig. 2D) does not show a specific pattern, revealing no relationship between discrepancies and values. Additionally, the Spearman's rank correlation coefficient is 0.084, confirming formally no correlation ($p = 0.40$). The 95% limits of agreement were − 6.51 and 6.58 (Fig. 2D).

Discussion

In this cross-sectional study with digestive cancer patients, we made two predictive models to estimate fat mass and muscle mass with high accuracy. These models were developed applying current machine learning methods throughout the whole process, from variable selection to model building. Additionally, these models used clinical variables which can easily be obtained during daily clinical practice. Accordingly, it is the first article in this field with this kind of approach.

Machine learning assembles a wide range of tools which could be used to explore data, to find patterns on patients' characteristics, to understand relationships between variables or to predict an outcome [30]. The implementation of these techniques to biomedical research has been growing during the last decade, however, they are not widely known and used [40]. Thus, we added a summary, in supplementary material, of some concepts of machine learning to bring this knowledge closer. In this study, a variety of methods was applied to analyze their usefulness and accuracy in the body composition field. Variable selection is one of the first problems to handle when a model is built. While manual variable selection is strongly influenced by our knowledge, techniques, as best subset selection or lasso, apply an algorithm and show an output as a measure of their performance to easily find the best set of variables. Besides, these techniques could show relevant variables not previously considered. Hence, this step becomes a part of the research process allowing to identify new predictors. Here, we found similar results during the selection process with different methods, reaching the best variable set for skeletal muscle with BSS and for fat mass with lasso regression. These comparable results, applying different processes as BSS and RF, highlight the importance of those variables. Additionally, these variable selection methods improved the accuracy of regression techniques. With regards to current predictive models, developed with anthropometric measurements, they are built with linear regression [25–29]. We showed how the incorporation of more flexible approaches can outperform linear models. Finally, cross-validation used for model selection, avoided overfitting as well as contributed to variable selection, allowing to find the best approach.

Skeletal muscle and fat mass have been considered central in several issues in cancer. Patients under oncology treatment and sarcopenia have more toxicity, worse clinical outcomes, and shorter survival [41–43]. Currently, the knowledge of skeletal muscle mass and fat mass can determine a bad prognosis in cancer patients and that may also be true in a wider patients population [41, 44].

Some aspects of our work can be considered. In this study, all patients showed weight loss because this cross-sectional study was nested in a prospective one. Although DXA is usually determined as the reference method, BIA is a valid and widespread methodology in patients without dehydration or edema. Concerning the methodology, we used LOOCV to test each model because the sample could not be split to have a test set due to its small size. Incomplete data is a common setting in health data sets. Leaving those as missing values implies dropping out the whole case when techniques as regression methods are used. Besides, this is a way to introduce bias. Thus, we used imputation methods as a way to obtain more accurate predictions.

This work shows the accuracy and utility of machine learning approaches to predict body composition. These findings prove the efficacy of these methods as well as the accuracy of our models, allowing the possibility of them being used as an investigation tool for pharmacokinetic models [45]. No general guidelines exist for drug adjustment in cachectic patients. Our models could be useful to adjust antineoplastic doses, in the clinical research and later in clinical practice.

It is now clear that body composition has an impact on oncology patients and their outcomes. Up to the present, no tool which allows determining body composition in daily practice without specific devices has been established. Working with current statistic learning techniques we were able to develop the first predictive model which uses daily practice clinical variables. Therefore, it can accurately predict fat mass and skeletal muscle mass. We believe that this could be a useful tool in the clinical setting allowing oncologists to obtain relevant information in an easy way, enabling more adequate patients' management and treatment.

Declarations

- **Funding**

The longitudinal study where data were collected (from August 2016 to January 2018) was conducted with a grant from the National Cancer Institute, Argentina (number 15001837). Then, the current study was developed without a grant support.

- **Conflicts of interest/Competing interests**

The authors declare that they have no conflict of interest.

- **Availability of data and material**

The datasets generated during and/or analyzed during the current study are available from the corresponding author on reasonable request.

- **Code availability**

R software was used for all data analysis. The code file used during the analysis is available on request from the corresponding author.

- **Authors' contributions**

Pablo Cresta Morgado: conceptualization, data curation, formal analysis, methodology, project administration, software, visualization, writing – original draft, and writing – review and editing.

Alfredo Hugo Navigante: conceptualization, data curation, investigation, resources, and writing – original draft

Adriana Pérez: conceptualization, formal analysis, methodology, project administration, software, supervision, validation, visualization, writing – original draft, and writing – review and editing.

- **Ethics approval**

The study was approved by The Ethics Board of the Gastroenterology National Hospital “Dr. Bonorino Udaondo” Buenos Aires, Argentina, and met the recommendations stated in the Helsinki Declaration.

- **Consent to participate**

Informed consent was obtained from all individual participants included in the study.

- **Consent for publication**

There are no conflicts regarding consent for publication.

References

1. Lightfoot A, McArdle A, Griffiths RD (2009) Muscle in defense. Crit Care Med 37:S384–S390 . https://doi.org/10.1097/CCM.0b013e3181b6f8a5
2. Porporato PE (2016) Understanding cachexia as a cancer metabolism syndrome. Oncogenesis 5:e200 . https://doi.org/10.1038/oncsis.2016.3
3. Hilmi M, Jouinot A, Burns R, Pigneur F, Mounier R, Gondin J, Neuzillet C, Goldwasser F (2019) Body composition and sarcopenia: the next-generation of personalized oncology and pharmacology? Pharmacol Ther 135–159 . https://doi.org/10.1016/j.pharmthera.2018.12.003
4. Trestini I, Carbognin L, Monteverdi S, Zanelli S, De Toma A, Bonaiuto C, Nortilli R, Fiorio E, Pilotto S, Di Maio M, Gasbarrini A, Scambia G, Tortora G, Bria E (2018) Clinical implication of changes in body composition and weight in patients with early-stage and metastatic breast cancer. Crit Rev Oncol Hematol 129:54–66 . https://doi.org/10.1016/j.critrevonc.2018.06.011
5. Shachar SS, Williams GR, Muss HB, Nishijima TF (2016) Prognostic value of sarcopenia in adults with solid tumours: A meta-analysis and systematic review. Eur J Cancer 57:58–67 . https://doi.org/10.1016/j.ejca.2015.12.030
6. Ida S, Watanabe M, Yoshida N, Baba Y, Umezaki N, Harada K, Karashima R, Imamura Y, Iwagami S, Baba H (2015) Sarcopenia is a Predictor of Postoperative Respiratory Complications in Patients with Esophageal Cancer. Ann Surg Oncol 22:4432–4437 . https://doi.org/10.1245/s10434-015-4559-3

7. Takagi K, Yoshida R, Yagi T, Umeda Y, Nobuoka D, Kuise T, Fujiwara T (2017) Radiographic sarcopenia predicts postoperative infectious complications in patients undergoing pancreaticoduodenectomy. BMC Surg 17:1–7 . https://doi.org/10.1186/s12893-017-0261-7
8. Tsaousi G, Kokkota S, Papakostas P, Stavrou G, Doumaki E, Kotzampassi K (2017) Body composition analysis for discrimination of prolonged hospital stay in colorectal cancer surgery patients. Eur J Cancer Care (Engl) 26:e12491 . https://doi.org/10.1111/ecc.12491
9. Malietzis G, Currie AC, Athanasiou T, Johns N, Anyamene N, Glynne-Jones R, Kennedy RH, Fearon KCH, Jenkins JT (2016) Influence of body composition profile on outcomes following colorectal cancer surgery. Br J Surg 103:572–580 . https://doi.org/10.1002/bjs.10075
10. Beumer JH, Chu E, Salamone SJ (2012) Body-Surface Area–Based Chemotherapy Dosing: Appropriate in the 21st Century? J Clin Oncol 30:3896–3897 . https://doi.org/10.1200/jco.2012.44.2863
11. Ratain M (1998) Body-surface area as a basis for dosing of anti-cancer agents: Science, Myth, or Habit? J Clin Oncol 16:2297–2298
12. Hopkins JJ, Sawyer MB (2017) A review of body composition and pharmacokinetics in oncology. Expert Rev Clin Pharmacol 10:947–956 . https://doi.org/10.1080/17512433.2017.1347503
13. Prado CMM (2013) Body composition in chemotherapy: The promising role of CT scans. Curr Opin Clin Nutr Metab Care 16:525–533 . https://doi.org/10.1097/MCO.0b013e328363bcfb
14. Andreoli A, Garaci F, Cafarelli FP, Guglielmi G (2016) Body composition in clinical practice. Eur J Radiol 85:1461–1468 . https://doi.org/10.1016/j.ejrad.2016.02.005
15. Prado CMM, Lima ISF, Baracos VE, Bies RR, McCargar LJ, Reiman T, MacKey JR, Kuzma M, Damaraju VL, Sawyer MB (2011) An exploratory study of body composition as a determinant of epirubicin pharmacokinetics and toxicity. Cancer Chemother Pharmacol 67:93–101 . https://doi.org/10.1007/s00280-010-1288-y
16. Shachar SS, Deal AM, Weinberg M, Williams GR, Nyrop KA, Popuri K, Choi SK, Muss HB (2017) Body composition as a predictor of toxicity in patients receiving anthracycline and taxane-based chemotherapy for early-stage breast cancer. Clin Cancer Res 23:3537–3543 . https://doi.org/10.1158/1078-0432.CCR-16-2266
17. Di Sebastiano KM, Mourtzakis M (2012) A critical evaluation of body composition modalities used to assess adipose and skeletal muscle tissue in cancer. Appl Physiol Nutr Metab 37:811–821 . https://doi.org/10.1139/h2012-079
18. Shen W, Punyanitya M, Wang Z, Gallagher D, St.-Onge M-P, Albu J, Heymsfield SB, Heshka S (2004) Visceral adipose tissue: relations between single-slice areas and total volume. Am J Clin Nutr 80:271–278
19. Shen W, Punyanitya M, Wang Z, Gallagher D, St.-Onge M-P, Albu J, Heymsfield SB, Heshka S (2004) Total body skeletal muscle and adipose tissue volumes: estimation from a single abdominal cross-sectional image. J Appl Physiol 97:2333–2338 . https://doi.org/10.1152/japplphysiol.00744.2004
20. Kazemi-Bajestani SMR, Mazurak VC, Baracos V (2016) Computed tomography-defined muscle and fat wasting are associated with cancer clinical outcomes. Semin Cell Dev Biol 54:2–10 . https://doi.org/10.1016/j.semcdb.2015.09.001
21. Prado CMM, Birdsell LA, Baracos VE (2009) The emerging role of computerized tomography in assessing cancer cachexia. Curr Opin Support Palliat Care 3:269–275 . https://doi.org/10.1097/SPC.0b013e328331124a
22. Pietrobelli A, Formica C, Wang Z, Heymsfield SB (1996) Dual-energy X-ray absorptiometry body composition model: review of physical concepts. Am J Physiol Metab 271:E941–E951 . https://doi.org/10.1152/ajpendo.1996.271.6.e941
23. Kyle UG, Bosaeus I, De Lorenzo AD, Deurenberg P, Elia M, Gómez JM, Heitmann BL, Kent-Smith L, Melchior JC, Pirlich M, Scharfetter H, Schols AMWJ, Pichard C (2004) Bioelectrical impedance analysis - Part I: Review of principles and methods. Clin Nutr 23:1226–1243 . https://doi.org/10.1016/j.clnu.2004.06.004
24. Wang J, Thornton JC, Kolesnik S, Pierson RN (2000) Anthropometry in body composition. An Overview. Ann N Y Acad Sci 904:317–326 . https://doi.org/10.1159/000191265
25. Truesdale KP, Roberts A, Cai J, Berge JM, Stevens J (2016) Comparison of Eight Equations That Predict Percent Body Fat Using Skinfolds in American Youth. Child Obes 12:314–323 . https://doi.org/10.1089/chi.2015.0020

26. Simões M, Severo M, Oliveira A, Ferreira I, Lopes C (2016) Predictive equations for estimating regional body composition: a validation study using DXA as criterion and associations with cardiometabolic risk factors. Ann Hum Biol 43:219–228 . https://doi.org/10.3109/03014460.2015.1054427
27. Aristizabal JC, Estrada-Restrepo A, García AG (2018) Development and validation of anthropometric equations to estimate body composition in adult women. Colomb Med 49:154–159 . https://doi.org/10.25100/cm.v49i2.3643
28. Ramirez-Zea M, Torun B, Martorell R, Stein AD (2006) Anthropometric predictors of body fat as measured by hydrostatic weighing in Guatemalan adults. Am J Clin Nutr 83:795–802
29. Durnin JVGA, 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. Br J Nutr 32:77–97
30. Gareth J, Witten D, Hastie T, Tibshirani R (2017) An Introduction to Statistical Learning
31. Mitchell TM (1997) Machine Learning. McGraw-Hill Science/Engineering/Math
32. Breiman LEO (2001) Random Forests. Mach Learn 45:5–32
33. Little RJ, Agostino RD, Cohen ML, Dickersin K, Emerson SS, Farrar JT, Frangakis C, Hogan JW, Molenberghs G, Murphy SA, Neaton JD, Rotnitzky A, Scharfstein D, Shih WJ, Siegel JP, Stern H (2012) The Prevention and Treatment of Missing Data in Clinical Trials. N Engl J Med 367:1355–1360
34. Buuren S van, Groothuis-oudshoorn K (2011) mice : Multivariate Imputation by Chained Equations in R. J Stat Softw 45:1–67
35. Buuren S Van, Oudshoorn CGM (2000) Multivariate Imputation by Chained Equations. TNO Prev. Heal. 1–39
36. Luo W, Phung D, Tran T, Gupta S, Rana S, Karmakar C, Shilton A, Yearwood J, Dimitrova N, Ho TB, Venkatesh S, Berk M (2016) Guidelines for Developing and Reporting Machine Learning Predictive Models in Biomedical Research: A Multidisciplinary View. J Med INTERNET Res 18:e323 . https://doi.org/10.2196/jmir.5870
37. Fearon K, Strasser F, Anker SD, Bosaeus I, Bruera E, Fainsinger RL, Jatoi A, Loprinzi C, MacDonald N, Mantovani G, Davis M, Muscaritoli M, Ottery F, Radbruch L, Ravasco P, Walsh D, Wilcock A, Kaasa S, Baracos VE (2011) Definition and classification of cancer cachexia: An international consensus. Lancet Oncol 12:489–495 . https://doi.org/10.1016/S1470-2045(10)70218-7
38. Bland JM, Altman DG (1999) Measuring agreement in method comparison studies. Stat Methods Med Res 8:135–160
39. Buuren S van (2018) Flexible Imputation of Missing Data, Second Edi. Chapman & Hall/CRC
40. Rajkomar A, Dean J, Kohane I (2019) Machine Learning in Medicine. N Engl J Med 380:1347–1358 . https://doi.org/10.1056/NEJMra1814259
41. Prado CMM, Baracos VE, McCargar LJ, Reiman T, Mourtzakis M, Tonkin K, Mackey JR, Koski S, Pituskin E, Sawyer MB (2009) Sarcopenia as a determinant of chemotherapy toxicity and time to tumor progression in metastatic breast cancer patients receiving capecitabine treatment. Clin Cancer Res 15:2920–2926 . https://doi.org/10.1158/1078-0432.CCR-08-2242
42. Wheler J, Tsimberidou AM, Hong D, Naing A, Jackson T, Liu S, Feng L, Kurzrock R (2009) Survival of patients in a Phase 1 clinic. Cancer 115:1091–1099 . https://doi.org/10.1002/cncr.24018
43. Reyes-Gibby CC, Wu X, Spitz M, Kurzrock R, Fisch M, Bruera E, Shete S (2008) Molecular epidemiology, cancer-related symptoms, and cytokines pathway. Lancet Oncol 9:777–785 . https://doi.org/10.1016/S1470-2045(08)70197-9
44. Schmidt D, Salahudeen A (2007) The obesity-survival paradox in hemodialysis patients: Why do overweight hemodialysis patients live longer? Nutr Clin Pract 22:11–15
45. Parsons HA, Baracos VE, Dhillon N, Hong DS, Kurzrock R (2012) Body composition, symptoms, and survival in advanced cancer patients referred to a phase I service. PLoS One 7:e29330 . https://doi.org/10.1371/journal.pone.0029330

Figures

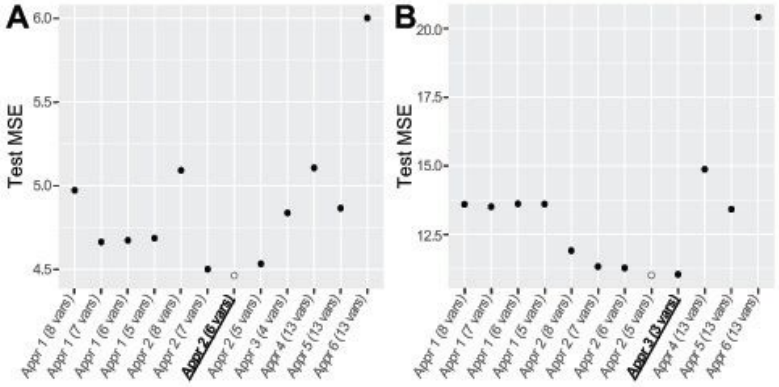

Figure 1

shows the results in terms of test MSE for skeletal muscle models (A) and for body fat mass models (B). As was mentioned, 6 different approaches for each were analyzed and for some of them variations with different numbers of variables were evaluated. White dots indicate the lowest test MSE and the label underlined in bold letters designates the model chosen. In every case the number of variables included in the model is indicated between brackets. Approach 1 (Appr 1): multicollinearity plus BSS plus multiple linear regression, 4 variable combinations were studied; Approach 2 (Appr 2): multicollinearity analysis plus best subset selection (BSS) plus generalized additive models (GAM), 4 different combinations of variables were analyzed; Approach 3 (Appr 3): multicollinearity plus lasso to perform variable selection plus GAM; Approach 4 (Appr 4): ridge regression; Approach 5 (Appr 5): lasso regression; Approach 6 (Appr 6): random forests.

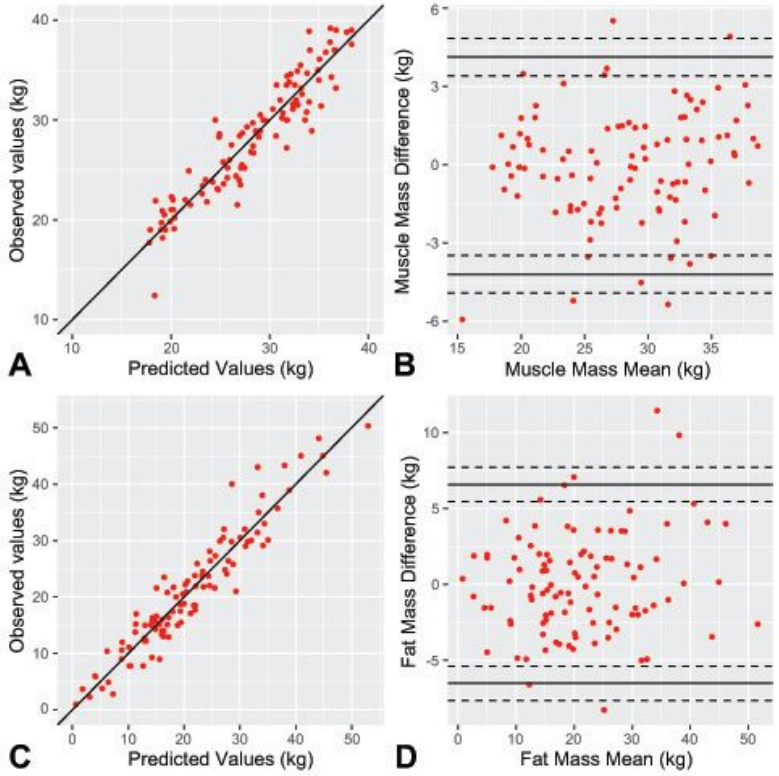

Figure 2

2A and 2B show skeletal muscle mass results. 2A is the scatter plot of skeletal muscle mass and shows observed values (measured during the study) and predicted values. The diagonal line indicates the equality line where there is maximal agreement. 2B displays the classical Bland and Altman plot in which is shows the difference between the measurements by the two methods (observed – predicted) for each subject against their mean. The horizontal solid lines represent the 95% limits of agreement and the dashed ones are the 95% confidence interval (CI) for each agreement limit. This plot allows to analyze the level of agreement

between methods as well as the existence of a relationship with the value (determined by the mean) and the discrepancies between methods (determined by the difference: observed – predicted). 2C and 2D show body fat mass results as was shown for skeletal muscle mass. 2C is the scatter plot of fat mass with observed and predicted values.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- ChecklistEQUATORNETWORK.pdf
- SupplementaryMaterial.docx