Multisite Technical and Clinical Performance Evaluation of Quantitative Imaging Biomarkers from 3D FDG PET Segmentations of Head and Neck Cancer Images

Brian J. Smith[1], John M. Buatti[3], Christian Bauer[2], Ethan J. Ulrich[2,4], Payam Ahmadvand[5], Mikalai M. Budzevich[6], Robert J. Gillies[6], Dmitry Goldgof[7], Milan Grkovski[8], Ghassan Hamarneh[5], Paul E. Kinahan[10], John P. Muzi[10], Mark Muzi[10], Charles M. Laymon[11,12], James M. Mountz[12], Sadek Nehmeh[13], Matthew J. Oborski[11], Binsheng Zhao[9], John J. Sunderland[14], and Reinhard R. Beichel[2]

[1]Departments of Biostatistics; [2]Electrical and Computer Engineering; [3]Radiation Oncology; and [4]Biomedical Engineering, The University of Iowa, Iowa City, IA; [5]School of Computing Science, Simon Fraser University, Burnaby, Canada; [6]H. Lee Moffitt Cancer Center & Research Institute, Department of Cancer Physiology, FL; [7]Department of Computer Science and Engineering, University of South Florida, FL; [8]Department of Medical Physics, Memorial Sloan Kettering Cancer Center, New York, NY; [9]Department of Radiology, Columbia University Medical Center, New York, NY; [10]Department of Radiology, The University of Washington Medical Center, Seattle, WA; [11]Department of Bioengineering, University of Pittsburgh, Pittsburgh, PA; [12]Department of Radiology, University of Pittsburgh, Pittsburgh, PA; [13]Department of Radiology, Weill Cornell Medical College, NY; and [14]Department of Radiology, The University of Iowa, Iowa City, IA

Corresponding Author:
Brian J. Smith, PhD
145 N Riverside Dr, Iowa City, IA, 52252-2007, phone: 319-384-1587;
E-mail: brian-j-smith@uiowa.edu

Key Words: FDG PET, head and neck cancer, multi-site performance analysis, radiomics, segmentation

Abbreviations: Quantitative imaging biomarkers (QIBs), fluorodeoxyglucose positron emission tomography (FDG PET), metabolic tumor volume (MTV), intraclass correlation coefficient (ICC), coefficient of total variation (wCV), odds ratio (OR), root mean square error (RMSE), rim average (RA), root mean square (RMS), standardized uptake value (SUV)

ABSTRACT

Quantitative imaging biomarkers (QIBs) provide medical image–derived intensity, texture, shape, and size features that may help characterize cancerous tumors and predict clinical outcomes. Successful clinical translation of QIBs depends on the robustness of their measurements. Biomarkers derived from positron emission tomography images are prone to measurement errors owing to differences in image processing factors such as the tumor segmentation method used to define volumes of interest over which to calculate QIBs. We illustrate a new Bayesian statistical approach to characterize the robustness of QIBs to different processing factors. Study data consist of 22 QIBs measured on 47 head and neck tumors in 10 positron emission tomography/computed tomography scans segmented manually and with semiautomated methods used by 7 institutional members of the NCI Quantitative Imaging Network. QIB performance is estimated and compared across institutions with respect to measurement errors and power to recover statistical associations with clinical outcomes. Analysis findings summarize the performance impact of different segmentation methods used by Quantitative Imaging Network members. Robustness of some advanced biomarkers was found to be similar to conventional markers, such as maximum standardized uptake value. Such similarities support current pursuits to better characterize disease and predict outcomes by developing QIBs that use more imaging information and are robust to different processing factors. Nevertheless, to ensure reproducibility of QIB measurements and measures of association with clinical outcomes, errors owing to segmentation methods need to be reduced.

INTRODUCTION

Quantitative imaging biomarkers (QIBs) provide medical image–derived intensity, texture, shape, and size features that have potential use in the characterization of disease and prediction of clinical outcomes. In the evolving field of *radiomics*, large numbers of potentially informative novel and diverse QIBs are extracted and studied for the personalization of disease treatment, particularly in oncology (1, 2). Examples of single

ISSN 2379-1381 https://doi.org/10.18383/j.tom.2020.00004

institution–based studies of imaging biomarkers include brain cancer (3, 4), head and neck cancer (5–8), lung cancer (9–13), nasopharyngeal carcinoma (14), prostate cancer (15, 16), and sarcoma (17). Other research has focused on performance of QIBs across multiple institutions, such as the analysis provided by Castelli et al. (18) regarding the predictive value of quantitative fluorodeoxyglucose positron emission tomography (FDG PET) in 45 studies of head and neck cancer.

Despite the growing body of radiomics research and the established use of some imaging biomarkers, such as metabolic tumor volume (MTV), few new QIBs have been adopted for clinical decision-making. Cancer Research UK and the European Organisation for Research and Treatment of Cancer, with NCI involvement, recently convened a consensus group to make recommendations for accelerating the clinical translation of imaging biomarkers. To that end, the group published a roadmap for navigating 3 main domains through which biomarker development passes: 1) discovery, 2) validation, and 3) qualification (19). In general, discovery is the process of identifying biomarkers associated with a disease or disease outcome of interest in a limited patient population; whereas, validation and qualification are formal assessments of biomarker performance and clinical utility in a broader population. Biomarker validation can be further divided into 2 complementary tasks, namely, *technical validation* and *clinical validation*, which focus on the quality of measured biomarker values and measured associations with disease, respectively. A third, qualification domain involves establishment of the fitness of biomarkers for specific clinical applications. Application of appropriate statistical methods is essential for the development of new clinically applicable QIBs. In particular, this process requires proper statistical estimation of measurement accuracy and precision for each of technical and clinical validation and proper statistical design and analysis of clinical trials for establishment of clinical utility.

In this paper, we use a new statistical approach for technical and clinical validation of QIBs derived from head and neck cancer FDG PET scans to investigate the impact of tumor segmentation variability across multiple institutions on the estimation of study power to design clinical trials (20). The approach uses a hierarchical Bayesian model to estimate systematic and random QIB measurement errors and simultaneously estimate the effects of these errors on study power to predict clinical outcomes. Specifically, our study is focused on 22 radiomic QIBs that were previously investigated regarding their ability to predict outcome in the treatment of head and neck cancer (21). The QIBs are derived from lesion segmentation resulting from an FDG PET/CT segmentation challenge involving 7 institutional members of the NCI Quantitative Imaging Network (QIN) (22, 23). All participating QIN members routinely use different approaches for lesion segmentation. Thus, the network provides an ideal setting within which to study the impact of segmentations on radiomic QIBs across methods and institutions. While our work focuses on errors because of using different segmentation tools, the used statistical methods are broadly applicable to other settings in which scanner, operator, or other image source differences contribute to QIB measurement errors.

Application to FDG PET imaging is of substantial interest for QIB development, because it is an established imaging approach for the quantification of cancer tumor burden (24–26). QIB extraction from FDG PET images involves several steps, including image acquisition and reconstruction. In addition, for many QIBs, segmentation of all tumors is required for calculating QIB values. Tumors may be segmented in a number of ways. Standard clinical practice is manual segmentation by trained experts (eg, radiation oncologists). Alternatively, a number of segmentation tools have been developed to help decrease human effort and increase segmentation consistency. These tools range from being semiautomated to fully automated (27). Although QIBs derived from tumor segmentations can be profoundly impacted by variation and bias in segmentation methods, existing studies provide little insight into the impacts of different methods on derived QIBs. In this work, we study this relevant issue.

Errors in PET-derived QIBs have been studied previously, primarily in terms of repeatability and reproducibility. Traverso et al. (28) performed a systematic review of 41 full-text articles to assess consensus regarding the robustness of commonly utilized radiomics QIBs for PET, CT, and MRI. The authors encountered error metric reporting of intraclass correlation coefficient (ICC) in 14 studies, correlation coefficient in 12, and various other descriptive statistics in 9. Bailly et al. (29) assessed variability of QIBs in relation to their dependence on different PET/CT reconstruction methods with coefficient of variation and percent deviation. Dice coefficient, ICC, and confidence interval half widths were used by Altazi et al. (30) to evaluate PET CT radiomic features in patients with cervical cancer. Kalpathy-Cramer et al. (31) report concordance correlation coefficients for the assessment of radiomic features from lung nodules in a multi-institutional study. Lu et al. (32) summarized reliability of radiomic features across image acquisition settings with R^2. Although the aforementioned studies use univariate or ANOVA-based statistics to estimate error, there are very few examples of simultaneous estimation of systematic and random errors. Beichel et al. (33) did use linear mixed effects regression to compare quality and variability of tumor volume measurements from the same QIN PET segmentation challenge analyzed herein. However, the Bayesian approach used in this study more generally analyzes the impact of all segmentation approaches simultaneously, provides estimates of study power, and includes 22 radiomic features and, therefore, illustrates a new statistical approach for QIB validation and qualification.

Our Bayesian statistical approach and QIN challenge application are described in the following section. Thereafter, analysis results are given to offer comparisons of measurements from challenge participants. Finally, a discussion is provided of the results and their implications for the current and future state of radiomic biomarker assessment and development.

METHODOLOGY

Quantitative Imaging Biomarkers

QIBs were derived from FDG PET/CT scans of patients with head and neck squamous cell carcinoma acquired at The University of Iowa Hospitals and Clinics (UIHC). Scans were collected, curated, and uploaded to TCIA (34) [collection: QIN-HEADNECK (35)] as part of the NCI QIN (22). A QIN segmentation challenge was

Table 1. Methods Used to Segment Tumors and Derive Quantitative Imaging Biomarkers in the QIN Segmentation Challenge

Method	Description	Operator(s)
Manual	Manual Segmentation	3 Radiation Oncologists
1	In-house software based on active contour segmentation	PhD research scientist
2	In-house software using a graph-based optimized segmentation	Radiation oncologist
3	Commercial software package Mirada Medical RTx	Imaging physicist
4	Combination of commercial software packages VCAR and PMOD	Medical physics postdoc
5	Commercial software package MIM	Imaging physicist
6	Commercial software package PMOD	Image analyst
7	In-house software based on 3D level-set segmentation	Medical image analysis graduate student

conducted in which a subset of 10 diverse pretreatment scans containing 47 lesions were segmented manually by 3 experienced radiation oncologists at the UIHC and by the following QIN sites: Columbia University Medical Center, H. Lee Moffitt Cancer Center and University of South Florida, Memorial Sloan Kettering Cancer Center, Simon Fraser University (Canada), University of Pittsburgh, The University of Iowa, and The University of Washington Medical Center. Sites were allowed to use segmentation tools of their choosing. Tools included both commercially available software and academic, in-house-developed segmentation algorithms. Deidentified summaries of the methods are given in Table 1. Further details of the challenge scanner acquisition and segmentation methods as well as evaluations of segmentation performance are given by Beichel et al. (33).

Forty-seven head and neck tumors in the 10 PET/CT scans were segmented using 7 different methods by the challenge participants. Each scan was segmented twice with a time interval

Table 2. Descriptions of the Quantitative Imaging Biomarkers Compared in the QIN Segmentation Challenge

QIB	Description (Unit)	Type
Max	Maximum value in region of interest (SUV)	C
Peak	Maximum average gray value that is calculated from a 1 cm^3 sphere placed within the region of interest (45) (SUV)	C
Mean	Mean value in region of interest (SUV)	C
MTV	Volume of region of interest (mL)	C
TLG	Total lesion glycolysis (mL)	C
Min	Minimum value in region of interest (SUV)	I
Standard	Standard deviation in region of interest (SUV)	I
RMS	Root mean square value in region of interest (SUV)	I
First Quartile	25th percentile value in region of interest (SUV)	I
Median	50th percentile value in region of interest (SUV)	I
Third Quartile	75th percentile value in region of interest (SUV)	I
Upper Adjacent	First value in region of interest not greater than 1.5 times the interquartile range (SUV)	I
Q1 Distribution	Percent of gray values that fall within the first quarter of the grayscale range within the region of interest (%)	I
Q2 Distribution	Percent of gray values that fall within the second quarter (%)	I
Q3 Distribution	Percent of gray values that fall within the third quarter (%)	I
Q4 Distribution	Percent of gray values that fall within the fourth quarter (%)	I
Glycolysis Q1	Lesion glycolysis calculated from the first quarter of the grayscale range within the region of interest (mL)	I
Glycolysis Q2	Lesion glycolysis calculated from the second quarter (mL)	I
Glycolysis Q3	Lesion glycolysis calculated from the third quarter (mL)	I
Glycolysis Q4	Lesion glycolysis calculated from the fourth quarter (mL)	I
SAM	Standardized added metabolic activity (46) (mL)	I
RA	Rim average; mean of uptake in a 2-voxel-wide rim region around region of interest (SUV)	I

Abbreviations: C, common clinical biomarkers; I, biomarkers provided by the 3D Slicer PET-IndiC extension.

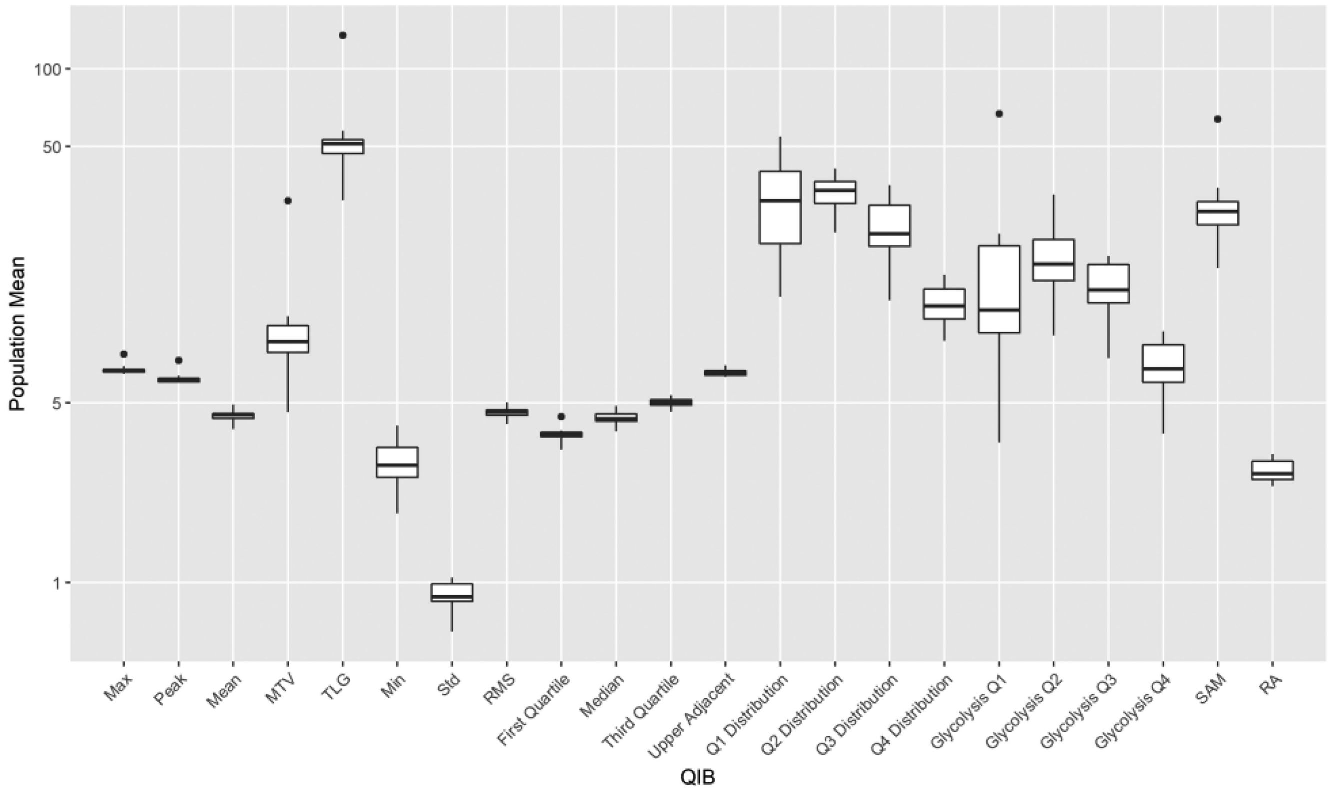


Figure 1. Boxplots showing the distribution of quantitative imaging biomarkers' (QIB) means calculated for each of the 8 segmentation methods for 47 lesions.

between initial and repeat segmentation. A challenge coordinator at The University of Iowa collected the segmentations and derived 22 QIBs with the 3D Slicer software for medical image informatics, image processing, and 3-dimensional visualization (36). The QIBs derived from lesion segmentations are summarized in Table 2 and include 5 of the most commonly used clinical biomarkers and 17 biomarkers available from the PET-IndiC extension (37) for the 3D Slicer, which were assessed in the context of outcome prediction by Beichel et al. (21). The PET-IndiC QIBs are generally designed to characterize standardized uptake value (SUV) patterns within segmented lesions by using descriptive statistics.

Statistical analysis focused on the quantification of random and systematic differences in QIB measurements across segmentation methods. For each method, descriptive means and standard deviations were computed on the population of segmented images. Agreement between and variability within the methods were estimated with a Bayesian regression modeling approach (20). This approach was taken to ensure that statistical inferences accounted for the study design, which included biomarkers derived from 8 different segmentation methods applied to a common set of 47 lesions, manual segmentation performed by 3 different operators, semiautomated segmentations performed by 1 operator each, 2 segmentations performed per operator and lesion. In brief, the statistical modeling of biomarker measurement $b_{i,j,k}$ for lesion i, operator j, and segmentation k is composed of the following series of mean and variance components:

$$b_{i,j,k} \sim N\left(\mu_{i,j}, \sigma^2_{\epsilon'}\right)$$

$$\mu_{i,j} \sim N\left(\mu_i, \sigma^2_{\epsilon''}\right)$$

$$\mu_i \sim N(\mu, \sigma^2_{\iota}).$$

Biomarker measurements from multiple readers and/or multiple segmentations of the same lesion are averaged together as μ_i. Within-lesion variance is $\sigma^2_{\epsilon} = \sigma^2_{\epsilon'} + \sigma^2_{\epsilon''}$ for the multiple readers and segmentations of manual segmentation and is $\sigma^2_{\epsilon} = \sigma^2_{\epsilon'}$ for the single reader and multiple segmentations of other methods. Between-lesion variance is σ^2_{ι}. The modeled means and variances are allowed to vary by segmentation method, denoted later with a subscript m. Thus, the application of the Bayesian model to the data provided estimates of mean differences between methods and differences between and within-lesion variability. Systematic mean differences were assessed relative to manual segmentation, the current standard for image segmentation of head and neck cancer. Systematic differences were estimated as the relative biases in population mean QIB measurements from semiautomated methods compared with manual segmentation:

$$Relative\ Bias_m = \frac{\mu_m - \mu_0}{\mu_0},$$

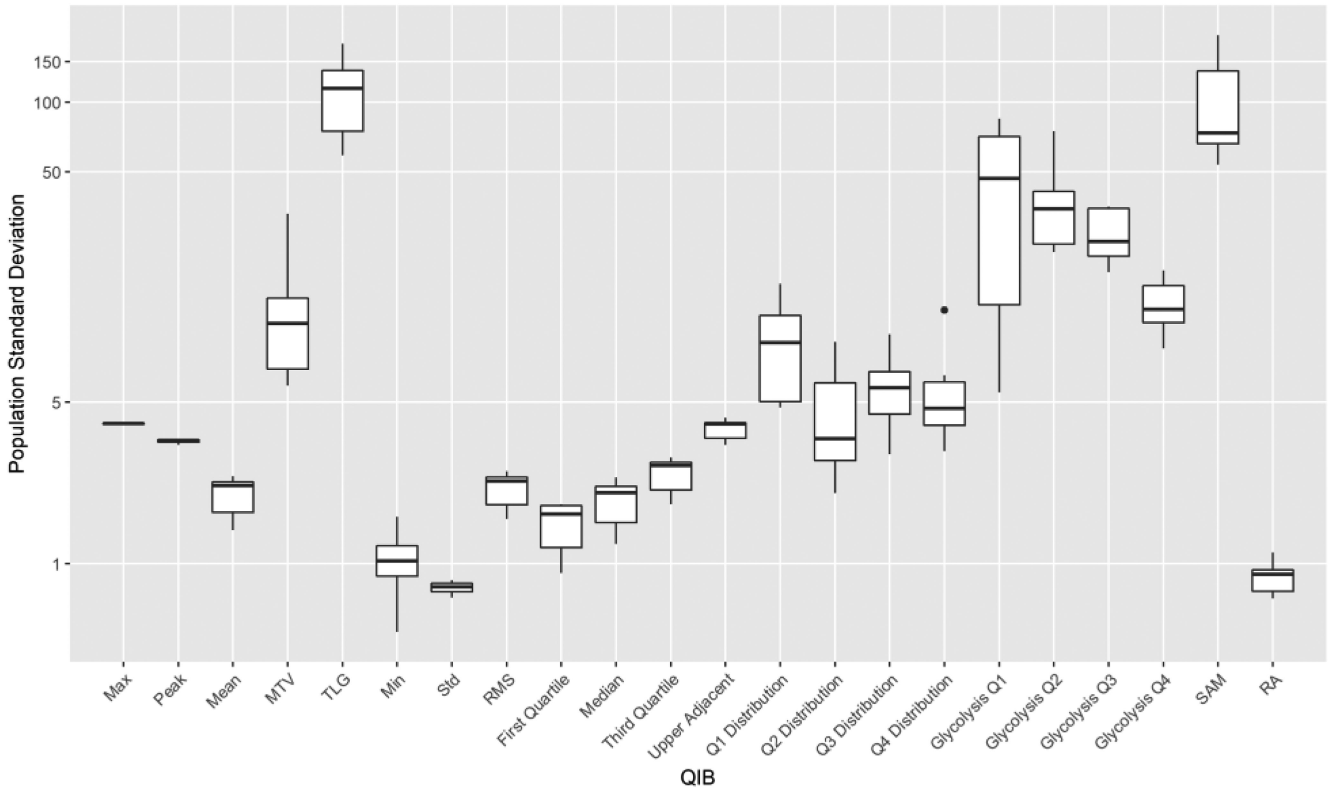


Figure 2. Boxplots showing the distribution of the QIB standard deviations calculated for each of the 8 segmentation methods for 47 lesions.

where μ_m is the biomarker mean for method m such that $m = 0$ is manual segmentation. Agreement was estimated with the concordance index (C-index) (38, 39). The C-index is a nonparametric, rank-based performance metric that can be interpreted as the probability a randomly selected pair of lesions will have QIB measurements with the same ordering on both segmentation methods being compared. Values of 1 and 0.5 represent perfect and chance concordance, respectively. Relative between-lesion variability was estimated with ICC and coefficient of total variation (wCV), respectively. ICC is defined as the variance in biomarker values between lesions relative to the total variance and is calculated as follows:

$$ICC_m = \frac{\sigma^2_{\iota m}}{\sigma^2_{\iota m} + \sigma^2_{\epsilon m}},$$

where $\sigma^2_{\iota_m}$ and $\sigma^2_{\epsilon_m}$ are between- and within-lesion variances for method m. Within-lesion variance is also known as *repeat error* and is the variability observed from repeated measurements on the same lesion. ICC values close to 1 indicate small repeat error relative to the total error. wCV is defined as total variability relative to the population mean and has the following form:

$$wCV_m = \frac{\sigma^2_{\iota m} + \sigma^2_{\epsilon m}}{\mu_m}.$$

Simulation studies were conducted to assess the impact of segmentation methods on estimating associations between QIBs and clinical outcomes, as described by Smith and Beichel (20). The general approach taken in the simulations is to define true relationships between manually segmented QIBs and a binary outcome and then to assess the degree to which QIBs from other segmentation methods can recover the true relationship. Specifically, probability (π) of a hypothetical binary outcome was defined in terms of a logistic relationship with manually segmented QIBs, such that,

$$\log\left(\frac{\pi_i}{1-\pi_i}\right) = \beta_0 + \beta_1 x_i,$$

where x_i is a QIB for lesion i. The β regression coefficients were chosen for the simulation study to reflect 50% prevalence of the outcome and an odds ratio (OR) of 2 for a 1 standard deviation increase in the manually segmented QIBs. Samples of 100 and 500 randomly selected lesions with their associated QIBs (x_i) were generated, outcome probabilities (π_i) calculated, and disease outcomes (y_i) simulated according the following Bernoulli probability distribution:

$$y_i \sim Bernoulli(\pi_i).$$

Then, logistic regression models were fit using the QIBs measured with other (semiautomated) methods to estimate statistical power to recover the true odds ratio (OR) of 2, a result useful for the determination of sample size in designing clinical trials. Also estimated from the simulations were method-specific biases

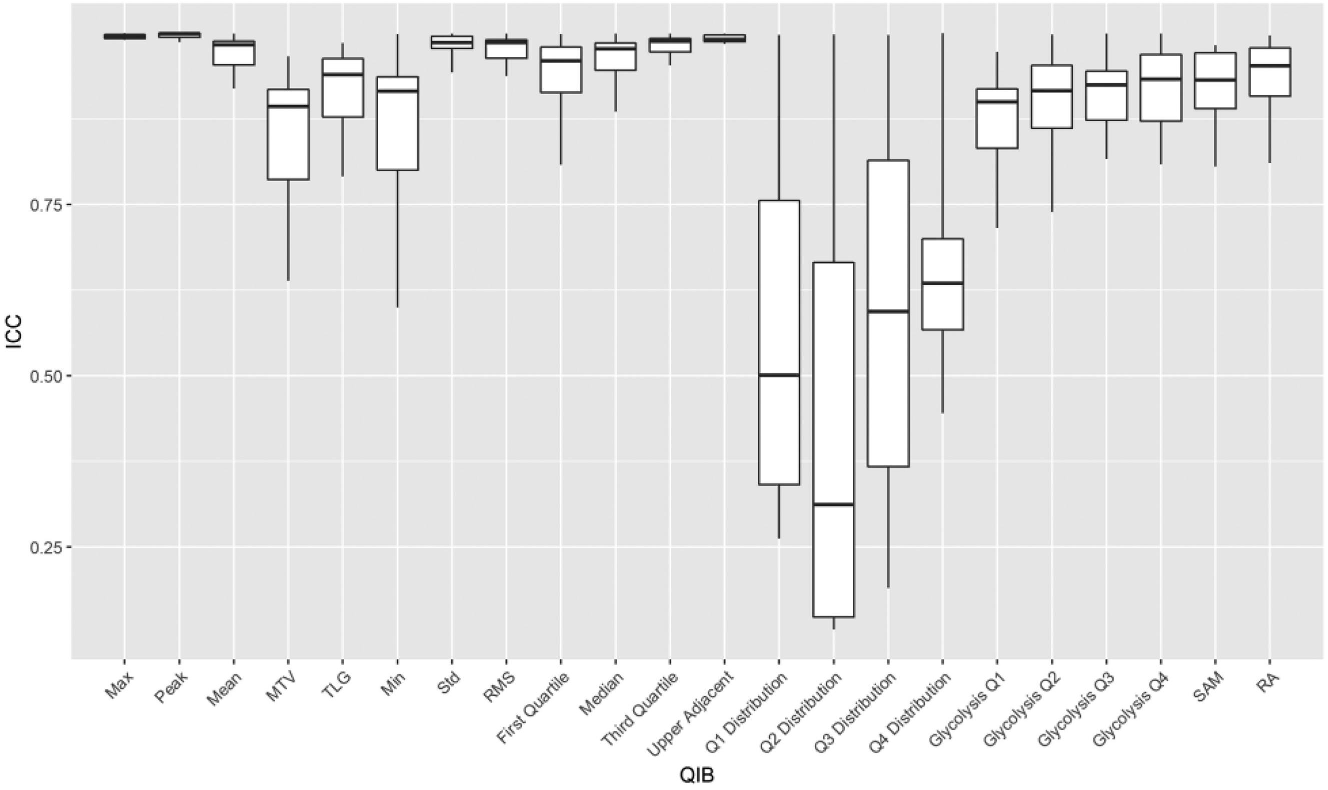


Figure 3. Boxplots showing the distribution of intraclass correlation coefficients (ICCs) across segmentation methods.

($Bias(OR_m) = OR_m - OR$), variances ($Var(OR_m)$), and root mean square error $\left(RMSE_m = \sqrt{Bias(OR_m)^2 + Var(OR_m)}\right)$ as a combination of estimation accuracy and precision. Lower bias, variance, and error indicate better estimation of the true OR value.

RESULTS

For descriptive comparisons, QIB means and standard deviations were computed over the measurements obtained from each segmentation method applied to the population of head and neck tumors included in the QIN challenge. The distributions of these method-specific population statistics are summarized with boxplots in Figures 1 and 2. In the plots, QIB values are displayed on the log scale to depict distributional variability relative to their different measurement scales. Distributions of the population means show how similar the methods are on average with respect to their QIB measurements. Accordingly, method means are most similar for the Max, Peak, and Mean clinical QIBs and similarly for the root mean square (RMS), First Quartile, Median, Third Quartile, and Upper Adjacent PET-IndiC QIBs. Similarities among the methods in the overall variability of their QIB measurements can be gauged by the distribution plots of population standard deviations. As with the population means, methods are most similar for the Max and Peak clinical QIBs. Otherwise, more dissimilarities are observed among the other clinical and PET-IndiC QIBs. Also noteworthy are the mean and standard deviation dissimilarities apparent in MTV measurements, indicating sensitivity of volumetric measurements to segmentation method. Method-specific estimates of the QIB means and standard deviations can be found in Supplemental Table 1.

Distributions of between- and within-method variability are summarized in Figures 3 and 4 with ICC and wCV, respectively. The ICC plots show the agreement of semiautomated segmentation methods with manual segmentation. Consistent with the population plots, there is near-perfect (ICC = 1) agreement among all methods for Max and Peak. High degrees of agreement are seen for Mean, Standard, RMS, First Quartile, Median, Third Quartile, and Upper Adjacent. Q1–Q4 Distributions exhibit very poor agreement; whereas the remaining QIBs have fairly good agreement. Within-method variability as measured by wCV tends to be low for many of the QIBs that have high agreement. Notable exceptions are MTV and total lesion glycolysis and several of the PET-IndiC QIBs that have moderate ICC but high wCV. Method-specific estimates of QIB variability as well as agreement are given in Supplemental Table 2.

Results of simulation studies are summarized in Figures 5 and 6 for N = 100 hypothetical binary clinical outcomes. As described in the methods, outcomes were repeatedly simulated based on $OR = 2$ for QIBs from manual segmentation. Bias, variability, and power were then calculated for ORs estimated with QIBs derived from the other semiautomated segmentation methods. As such manual segmentation defines the true relationship between QIBs and clinical outcomes, and the results quantify the quality of estimates that can be obtained with the other methods.

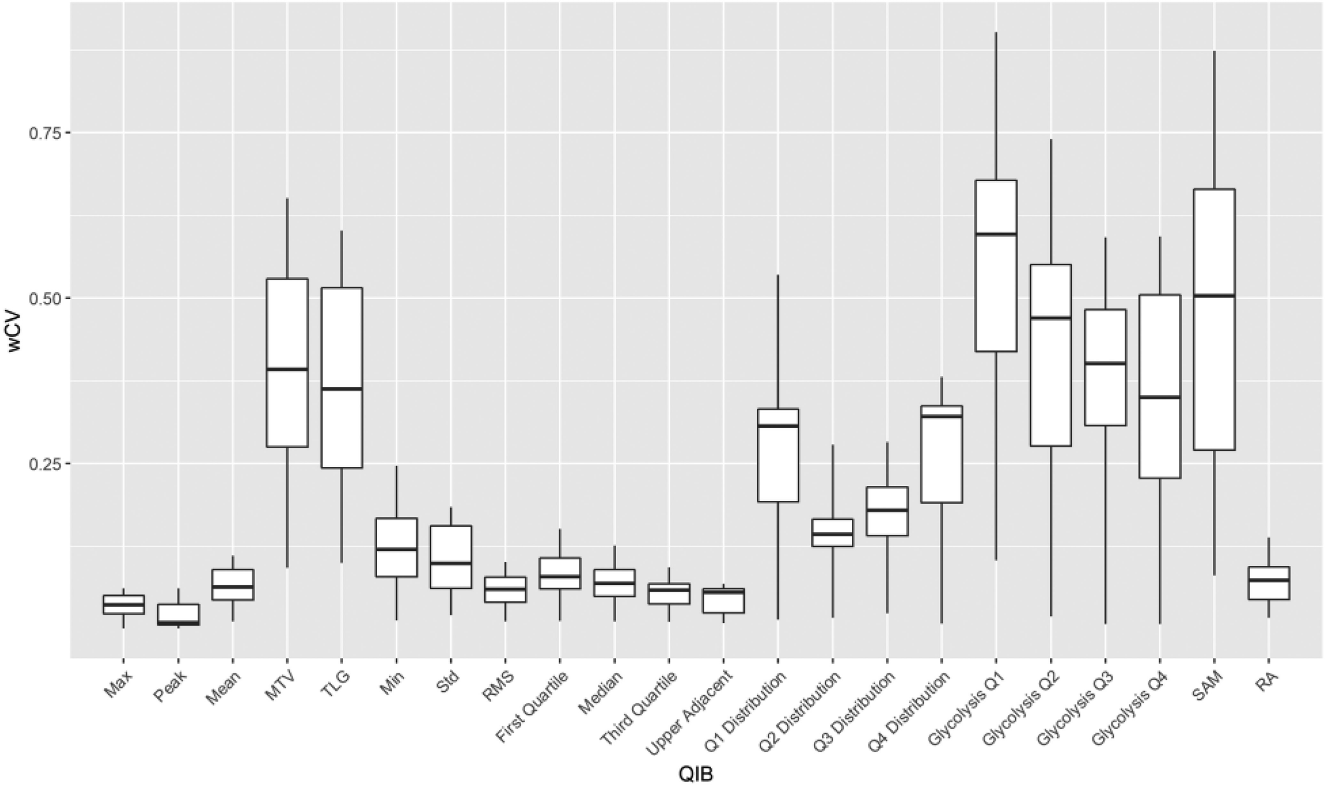


Figure 4. Boxplots showing the distribution of QIB within-method coefficients (wCVs) of variability across segmentation methods.

Taking into account both estimation bias and variability, RMSE values plotted in Figure 5, and tabulated in Supplemental Table 3, show relatively low error for Upper Adjacent, RMS, Third Quartile, Mean, Glycolysis Q4, Max, and Peak. Statistical power to detect effects of the QIBs, at the 5% level of significance, is summarized in the heatmap of Figure 6. The QIBs and methods in the heatmap are ordered according to similarity measures from hierarchical clustering of their powers. Dendrogram clustering of the 2 are displayed to the top and right of the heatmap. Power is generally inversely related to RMSE. Overall, the effects of clinical QIBs, compared to those of PET-IndiC QIBs, were less affected by the segmentation method. With respect to methods, QIB measurements from segmentation method 2 are most similar to those from manual segmentation. After that, the grouping of methods 3 and 7 are most similar to those of method 2. Method 4 produced the outlying values depicted as individual dots on the boxplots discussed previously and has the lowest power. Accordingly, a clinical trial planning to use segmentation method 4 would require a larger sample size for most of the QIBs. Likewise, within a method, the study power would vary depending on the biomarker for which a trial is being designed.

Based on the previously discussed measures of agreement, variability, and power, hierarchical clustering was used to identify QIB groupings for which the impact of segmentation was either low, moderate, high, or extreme. Table 3 presents the clustering results and summarizes performance measures aggregated over the 7 semiautomated segmentation methods. Coefficients of variation computed from the segmentation-specific population means were 7.8%, 7.3%, 53.8%, and 27.7% in the low, moderate, high, and extremely impacted biomarker groups, respectively. Agreements to manual segmentation as measured by absolute relative biases were 6.7%, 13.2%, 52.0%, and 51.1%. The extreme group stood apart from the other as having comparatively poor ICC of 0.603 and power of 26.9%. Average ICC for the low through highly impacted groups was markedly better at 0.993, 0.966, and 0.892, and powers were 85.1%, 78.1%, and 67.7%.

DISCUSSION

Segmentation Impact on QIBs

In this work, a unified Bayesian modeling approach was applied to estimate QIB measurement errors and their effects on statistical power. It enables quantification and comparison of the effects of different tumor segmentation methods on the panel of 22 QIBs. Clinical QIBs have long been used in clinical research and practice to characterize disease and to assess disease progression. A widely used example is the RECIST criteria for defining tumor response in clinical trials in terms of change in imaged tumor size (40, 41). Improvements in and increasing access to medical imaging have fueled interest in other, more advanced QIBs indicated in the panel of PET-IndiC QIBs. Unfortunately, QIB measurements are subject to errors from multiple sources, including scanner makes and models, settings, reconstruction algorithms,

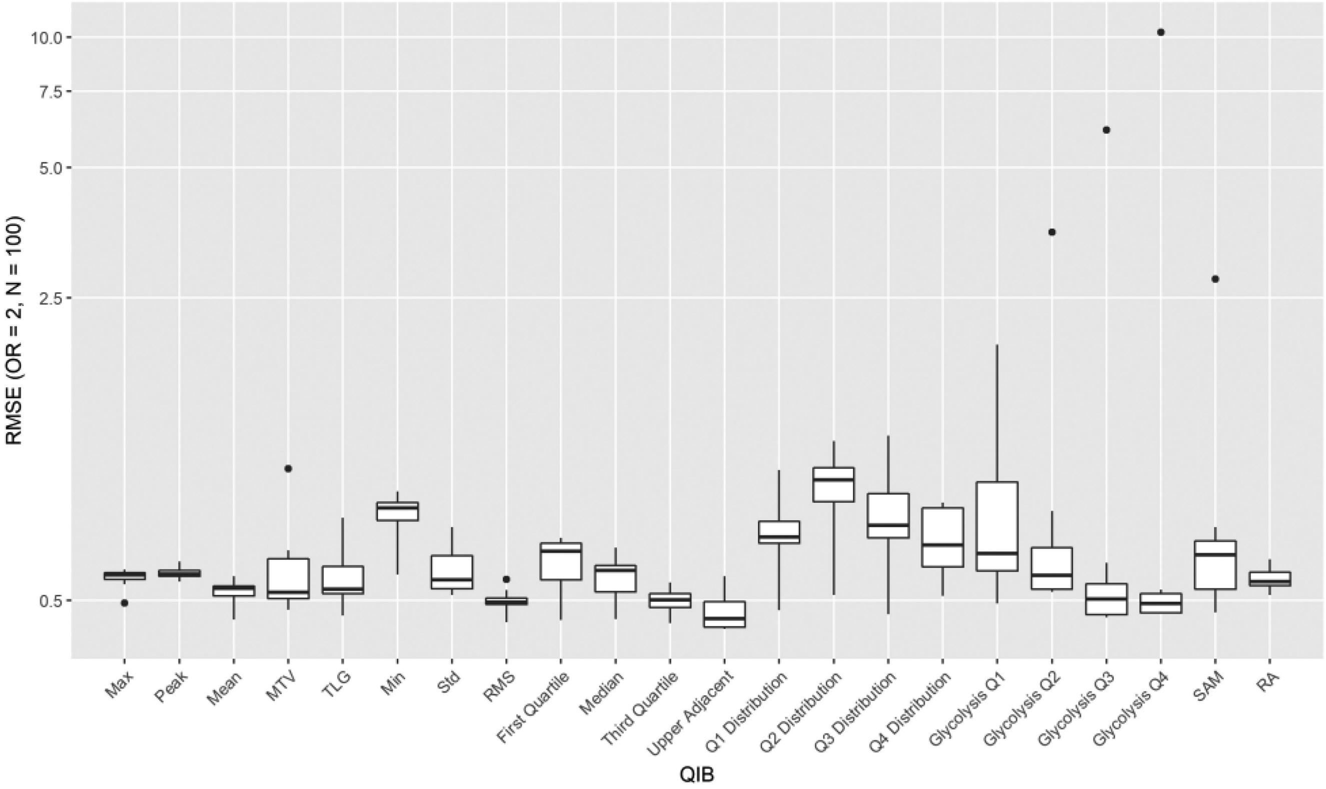


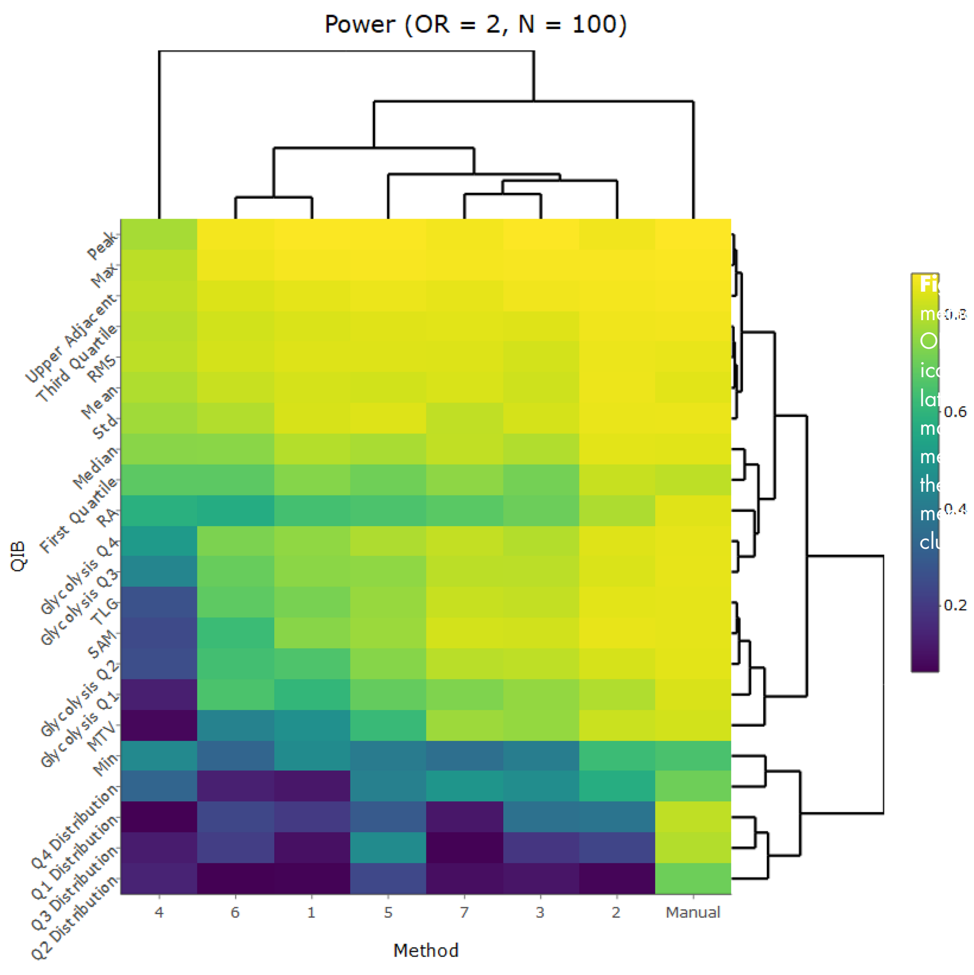
Figure 5. Boxplots of QIB root mean square error (RMSE) comparing method-specific odds ratios (ORs) estimated from hypothetical binary clinical outcomes simulated from QIB relationships defined by manual segmentations. RMSE is calculated as the square root of the estimated odds ratio bias squared plus its variance.

segmentation methods, and biologic variability. Our approach enables illustration of the effects of segmentation methods on random and systematic differences as well as statistical power.

Segmentation of tumors defines volumes of interest within which voxel intensities are used to calculate various QIBs that quantify different properties of tumors. Typically, manual and semiautomated tumor segmentation approaches—as used within this work—are subject to various degrees and types of variation in generated VOIs. An example for tumor segmentation differences between QIN sites is depicted in Figure 7. Intuition would suggest that segmentation methods have less of an effect on QIBs whose calculations are less dependent on accurate VOI definition. Indeed, several new segmentation methods have been motivated by the insensitivity of biomarkers extracted from them. For instance, Echegaray et al. (42, 43) propose “core samples” and “digital biopsy” segmentation methods for which several intensity and texture features were shown to be consistent with a reference standard. In our comparison of multiple segmentation methods, QIB quantile measures (Max, Peak, First Quartile, Median, Third Quartile, and Upper Adjacent) extracted primarily from interior voxel intensities have particularly high agreement of the population means and relatively high agreement of the population standard deviations, high ICC, low wCV, low RMSE, and high power. The measurements of mean-based QIBs (Mean, Standard, RMS, and rim average [RA]) also exhibit relatively high degrees of reliability. However, such relatively simple QIBs might not be able to capture desirable characteristics of tumors, such as texture. The remaining QIBs, which more broadly utilize the VOI for QIB calculation, are more affected by the segmentation method, but might provide relevant information. Thus, it is imperative to study the impact of tumor segmentation variability on subsequent predictive modeling. For example, to discover a relationship between QIB and outcome, more samples might be needed for a segmentation method that is more prone to segmentation variability than a method that is less prone to variability.

Consequently, our technical performance assessments were designed to assess impact on QIB measurements, because ultimate interest is often on QIB performance in the prediction of clinical outcomes, also known as *clinical performance*. To address clinical performance, our Bayesian approach provides simulation study results to characterize the effect of segmentation on the ability to recover associations between QIBs and a hypothetical clinical outcome. Many of the quantile measures that had good technical performance also had good clinical performance, that is, low RMSE and high power. A few exceptions were the lower power of Median, First Quartile, and RA. In addition, the low statistical power and high variation across segmentation methods for MTV are noteworthy because many studies propose to utilize MTV for outcome prediction.

Typically, segmentation methods are evaluated regarding only their segmentation performance. Our statistical analysis



Fi [illegible]
me [illegible] ecific powers to detect
O [illegible] ations between hypothet-
ic [illegible] y clinical outcomes simu-
lat [illegible] relationships defined by
m [illegible] egmentation. QIBs and
m [illegible] re ordered according to
th [illegible] ity of their powers as
m [illegible] by hierarchical
cl [illegible]

approach enables the selection of methods regarding their suitability for specific QIBs. Method 4 stands out as having noticeably lower power than the other methods. In general, power varies differentially across QIBs and methods, thus helping explain why a QIB may be identified as statistically significant in one research setting but not in another when different segmentation methods are used.

Implications

The illustrated statistical approach can aid QIB development by providing estimates of technical and clinical performance for biomarker validation and of statistical power for clinical trial design. The application considered involves development of QIBs derived from different semiautomated segmentation methods. Such computer-aided analysis of medical images has the potential to advance the development of QIBs by decreasing the time needed to extract them and by increasing the consistency of their measurements. Image analysis methods are advancing rapidly with several semiautomated tools currently available for the segmentation and quantification of FDG PET images. Given the range and freedom of choices that exist, understanding the effects of different tools on the technical and clinical performance of QIBs derived from them is essential. To that end, the technical and clinical performance analysis results provided by the present study represent a baseline and provide a starting point for future improvements in imaging biomarker quantification. Furthermore, our analysis explores performance within a multisite (QIN challenge) setting in which different segmentation tools are used. Results show degrees of systematic and random differences between sites that highlight the need for improved consistency of segmentation tool algorithms and their application. Multiple courses of action should be considered to improve consistency. Tool application guidelines and training are important at the user level. In addition, tool consistency could be improved with application-specific method development and benchmarking against publicly available and clinically relevant data sets.

Improved consistency of computer-aided tools will increase the utility of QIBs for disease characterization and response assessment. This is particularly relevant for multicenter clinical trials and the field of radiomics in general where images may be processed quantitatively by different operators and at different institutions. Future adoption of standards for tool development and statistical assessment as well as reduced requirements for user operability would benefit image analysis in such decentralized applications. The current state, however, is quite heterogeneous with respect to technologies, operators, and assessments.

Table 3. Summary of Performance Metrics for QIBs Grouped by Segmentation Impact

QIB by Segmentation Impact	Population Mean CV	Average Absolute Relative Bias	Average wCV	Average ICC	Average Power
Low					
Max	0.060	0.039	0.033	0.996	0.866
Peak	0.068	0.048	0.019	0.997	0.864
Standard	0.146	0.139	0.096	0.988	0.822
Upper Adjacent	0.039	0.041	0.042	0.993	0.854
Group Mean (SD)	0.078 (0.047)	0.067 (0.049)	0.048 (0.033)	0.993 (0.004)	0.851 (0.020)
Moderate					
Mean	0.063	0.143	0.061	0.975	0.829
RMS	0.058	0.126	0.057	0.980	0.839
First Quartile	0.085	0.176	0.078	0.947	0.727
Median	0.070	0.144	0.067	0.967	0.788
Third Quartile	0.049	0.098	0.054	0.984	0.841
RA	0.111	0.106	0.072	0.940	0.660
Group Mean (SD)	0.073 (0.022)	0.132 (0.029)	0.065 (0.009)	0.966 (0.018)	0.781 (0.074)
High					
MTV	0.559	0.370	0.367	0.910	0.703
TLG	1.054	1.542	0.528	0.861	0.623
Glycolysis Q1	0.380	0.333	0.414	0.891	0.677
Glycolysis Q2	0.269	0.248	0.371	0.910	0.726
Glycolysis Q3	0.284	0.254	0.341	0.920	0.747
Glycolysis Q4	0.479	0.392	0.454	0.915	0.700
SAM	0.559	0.370	0.367	0.910	0.703
Group Mean (SD)	0.538 (0.281)	0.52 (0.459)	0.409 (0.064)	0.892 (0.031)	0.677 (0.063)
Extreme					
Min	0.232	0.672	0.108	0.894	0.434
Q1 Distribution	0.459	1.191	0.268	0.521	0.237
Q2 Distribution	0.176	0.148	0.149	0.389	0.113
Q3 Distribution	0.318	0.339	0.180	0.556	0.198
Q4 Distribution	0.198	0.203	0.253	0.655	0.362
Group Mean (SD)	0.277 (0.115)	0.511 (0.431)	0.192 (0.068)	0.603 (0.188)	0.269 (0.129)

Abbreviations: CV, coefficient of variation; wCV, within coefficient of variation; ICC, intraclass correlation coefficient.

Limitations

The QIN challenge data analyzed in this study has some notable limitations. First, its scope is limited only to the effect of segmentation method on QIB measurements. Other factors such as scanner type, settings, and reconstruction algorithm will also affect the measurements. All images were obtained at the same institution so as to reduce the effects of image acquisition differences on results obtained in the QIN challenge. Second, the challenge results may not generalize to non-head and neck cancers, as stability of biomarker measurements has been observed to differ across cancer types (44). These 2 limitations are characteristic of the data source and not the statistical approach, which can be applied to estimate measurement error and predictive performance in other settings in which additional sources of measurement error are present. Third, there is no absolute ground truth segmentation for head and neck tumors. Instead, manual segmentation was used as a surrogate ground truth, or reference standard, for the calculation of agreement (C-index) and for the simulation studies to estimate RMSE and statistical power. To mitigate variability in this reference standard, manual segmentations were performed by 3 expert radiation oncologists at 2 separate time points and combined to derive reference QIB measurements for each tumor. Fourth, a synthetic simulation study was conducted to assess clinical performance rather than using actual clinical outcomes from patients. The advantage of this approach is that the true relationship between QIBs and simulated outcomes is known and can thus be used to estimate RMSE and power. Moreover, simulation is a valid and commonly used approach for the design of clinical trials. The disadvantage is that the statistical model used in the simulation may not fully reflect the complexities of true relationships between QIBs and clinical outcomes.

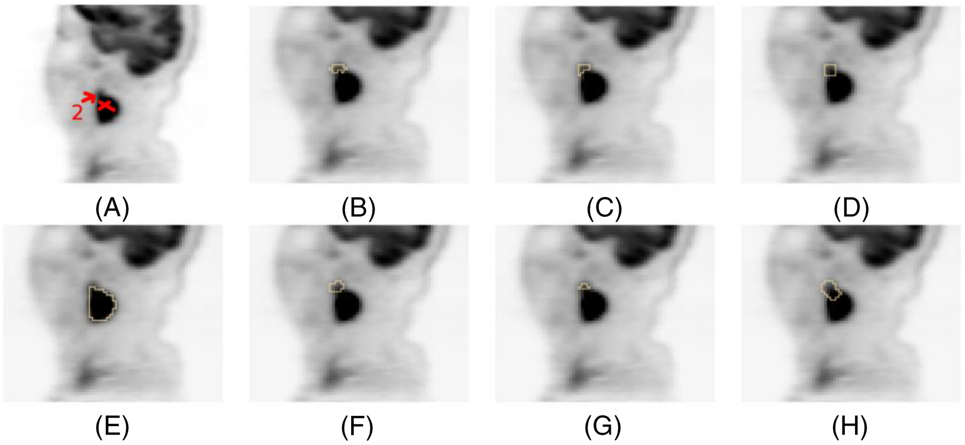


Figure 7. Example head and neck positron emission tomography (PET)/computed tomography (CT) segmentations of a cancerous lymph node. Guidance image provided to challenge participants indicating to segment a lesion (indicated as "2") located next to a large primary tumor (X), which should be excluded in the segmentation (A). Substantial differences in derived quantitative imaging biomarkers can result from segmentation methods that correctly distinguish the lesion (B, C, F and G) versus those that leak into the primary tumor (D and H) or fail to distinguish the lymph node from the primary tumor (E).

CONCLUSIONS

QIBs are becoming increasingly important in the characterization, treatment, and prognostication of disease. Clinical markers such as maximum SUV and tumor volume have a long history of use. The simplicity of their calculations lend themselves well to widespread adoption. However, that simplicity may limit their utility as prognostic indicators. Thus, there is interest in more advanced markers that utilize texture, shape, and intensity information from imaged tumors. Such features can be more prone to measurement errors owing to differences in segmentation methods or other image acquisition or processing steps. The used statistical approach can help quantify QIB measurement error in real-world (eg, multi-institutional) settings for which they are being developed. Results from the approach could be used to prioritize QIBs that are less sensitive to measurement error, to identify standardizations needed in the process by which QIBs are derived, or to determine statistical power for clinical trial design. For example, our finding that PET-IndiC features Standard, RMS, First/Third Quartile, Upper Adjacent, and RA have technical performance similar to maximum SUV and tumor volume suggest that these more advance markers can be measured as reliably and precisely as standard clinical makers. Over all of the markers analyzed, we observed a wide range of performances and thus conclude that errors due to segmentation methods need to be reduced. Therefore, we recommend establishment of reference imaging data set collections and reference segmentations against which segmentation methods can be benchmarked and tuned to ensure harmonization of QIBs. The presented findings summarize the current state of QIB variability and systematic differences owing to segmentation methods used by NCI QIN members. Moreover, the statistical analysis of technical and clinical QIB performance offers an approach that could be used in the future to develop QIBs in other disease and imaging settings.

Supplemental Materials

Supplemental Tables 1–3: https://doi.org/10.18383/j.tom.2020.00004.sup.01

ACKNOWLEDGMENTS

This research was funded in part by The University of Iowa Holden Comprehensive Cancer Center through the Cancer Center Support Grant P30 CA086862 and the National Institutes of Health Grants U01 CA140206, UL1TR002537, U24 CA180918, and R50 CA211270.

Disclosures: No disclosures to report.

Conflict of Interest: The authors have no conflict of interest to declare.

REFERENCES

1. Cook GJR, Siddique M, Taylor BP, Yip C, Chicklore S, Goh VJC, Imaging T. Radiomics in PET: principles and applications. Clin Transl Imaging. 2014;2:269–276.
2. Kumar V, Gu Y, Basu S, Berglund A, Eschrich SA, Schabath MB, Forster K, Aerts HJ, Dekker A, Fenstermacher D, Goldgof DB, Hall LO, Lambin P, Balagurunathan Y, Gatenby RA, Gillies RJ. Radiomics: the process and the challenges. Magn Reson Imaging. 2012;30:1234–1248.
3. Oborski MJ, Laymon CM, Lieberman FS, Qian Y, Drappatz J, Mountz JM. [(18)F]ML-10 PET: initial experience in glioblastoma multiforme therapy response assessment. Tomography. 2016;2:317–324.
4. Coolens C, Driscoll B, Foltz W, Pellow C, Menard C, Chung C. Comparison of voxel-wise tumor perfusion changes measured with Dynamic Contrast-Enhanced (DCE) MRI and volumetric DCE CT in patients with metastatic brain cancer treated with radiosurgery. Tomography. 2016;2:325–333.

5. Parmar C, Grossmann P, Rietveld D, Rietbergen MM, Lambin P, Aerts HJ. Radiomic machine-learning classifiers for prognostic biomarkers of head and neck cancer. Front Oncol. 2015;5:272.
6. Lambin P, Leijenaar RTH, Deist TM, Peerlings J, de Jong EEC, van Timmeren J, Sanduleanu S, Larue R, Even AJG, Jochems A, van Wijk Y, Woodruff H, van Soest J, Lustberg T, Roelofs E, van Elmpt W, Dekker A, Mottaghy FM, Wildberger JE, Walsh S. Radiomics: the bridge between medical imaging and personalized medicine. Nat Rev Clin Oncol. 2017;14:749–762.
7. You D, Aryal M, Samuels SE, Eisbruch A, Cao Y. Temporal feature extraction from DCE-MRI to identify poorly perfused subvolumes of tumors related to outcomes of radiation therapy in head and neck cancer. Tomography. 2016;2:341–352.
8. Ulrich EJ, Menda Y, Boles Ponto LL, Anderson CM, Smith BJ, Sunderland JJ, Graham MM, Buatti JM, Beichel RR. FLT PET radiomics for response prediction to chemoradiation herapy in head and neck squamous cell cancer. Tomography. 2019;5:161–169.
9. Yip SS, Kim J, Coroller TP, Parmar C, Velazquez ER, Huynh E, Mak RH, Aerts HJ. Associations between somatic mutations and metabolic imaging phenotypes in non-small cell lung cancer. J Nucl Med. 2017;58:569–576.
10. Patil R, Mahadevaiah G, Dekker A. An approach toward automatic classification of tumor histopathology of non-small cell lung cancer based on radiomic features. Tomography. 2016;2:374–377.
11. Paul R, Hawkins SH, Balagurunathan Y, Schabath MB, Gillies RJ, Hall LO, Goldgof DB. Deep feature transfer learning in combination with traditional features predicts survival among patients with lung adenocarcinoma. Tomography. 2016;2:388–395.
12. Mattonen SA, Davidzon GA, Bakr S, Echegaray S, Leung ANC, Vasanawala M, Horng G, Napel S, Nair VS. [18F] FDG Positron Emission Tomography (PET) tumor and penumbra imaging features predict recurrence in non-small cell lung cancer. Tomography. 2019;5:145–153.
13. Paul R, Schabath M, Balagurunathan Y, Liu Y, Li Q, Gillies R, Hall LO, Goldgof DB. Explaining deep features using radiologist-defined semantic features and traditional quantitative features. Tomography. 2019;5:192–200.
14. Zhang B, He X, Ouyang F, Gu D, Dong Y, Zhang L, Mo X, Huang W, Tian J, Zhang S. Radiomic machine-learning classifiers for prognostic biomarkers of advanced nasopharyngeal carcinoma. Cancer Lett. 2017;403:21–27.
15. Parra NA, Lu H, Choi J, Gage K, Pow-Sang J, Gillies RJ, Balagurunathan Y. Habitats in DCE-MRI to predict clinically significant prostate cancers. Tomography. 2019;5:68–76.
16. McGarry SD, Bukowy JD, Iczkowski KA, Unteriner JG, Duvnjak P, Lowman AK, Jacobsohn K, Hohenwalter M, Griffin MO, Barrington AW, Foss HE, Keuter T, Hurrell SL, See WA, Nevalainen MT, Banerjee A, LaViolette PS. Gleason probability maps: a radiomics tool for mapping prostate cancer likelihood in MRI space. Tomography. 2019;5:127–134.
17. Huang W, Beckett BR, Tudorica A, Meyer JM, Afzal A, Chen Y, Mansoor A, Hayden JB, Doung YC, Hung AY, Holtorf ML, Aston TJ, Ryan CW. Evaluation of soft tissue sarcoma response to preoperative chemoradiotherapy using dynamic contrast-enhanced magnetic resonance imaging. Tomography. 2016;2:308–316.
18. Castelli J, De Bari B, Depeursinge A, Simon A, Devillers A, Roman Jimenez G, Prior J, Ozsahin M, de Crevoisier R, Bourhis J. Overview of the predictive value of quantitative 18 FDG PET in head and neck cancer treated with chemoradiotherapy. Crit Rev Oncol Hematol. 2016;108:40–51.
19. O'Connor JPB, Aboagye EO, Adams JE, Aerts HJWL, Barrington SF, Beer AJ, Boellaard R, Bohndiek SE, Brady M, Brown G, Buckley DL, Chenevert TL, Clarke LP, Collette S, Cook GJ, deSouza NM, Dickson JC, Dive C, Evelhoch JL, Faivre-Finn C, Gallagher FA, Gilbert FJ, Gillies RJ, Goh V, Griffiths JR, Groves AM, Halligan S, Harris AL, Hawkes DJ, Hoekstra OS, Huang EP, Hutton BF, Jackson EF, Jayson GC, Jones A, Koh D-M, Lacombe D, Lambin P, Lassau N, Leach MO, Lee T-Y, Leen EL, Lewis JS, Liu Y, Lythgoe MF, Manoharan P, Maxwell RJ, Miles KA, Morgan B, Morris S, Ng T, Padhani AR, Parker GJM, Partridge M, Pathak AP, Peet AC, Punwani S, Reynolds AR, Robinson SP, Shankar LK, Sharma RA, Soloviev D, Stroobants S, Sullivan DC, Taylor SA, Tofts PS, Tozer GM, van Herk M, Walker-Samuel S, Wason J, Williams KJ, Workman P, Yankeelov TE, Brindle KM, McShane LM, Jackson A, Waterton JC. Imaging biomarker roadmap for cancer studies. Nat Rev Clin Oncol. 2017;14:169–186.
20. Smith BJ, Beichel RR. A Bayesian framework for performance assessment and comparison of imaging biomarker quantification methods. Stat Methods Med Res. 2019;28:1003–1008.
21. Beichel RR, Ulrich EJ, Smith BJ, Bauer C, Brown B, Casavant T, Sunderland JJ, Graham MM, Buatti JM. FDG PET based prediction of response in head and neck cancer treatment: assessment of new quantitative imaging features. PLoS One. 2019;14:e0215465.
22. National Cancer Institute. Quantitative Imaging Network. 2018 Available from: https://imaging.cancer.gov/programs_resources/specialized_initiatives/qin.
23. Farahani K, Kalpathy-Cramer J, Chenevert TL, Rubin DL, Sunderland JJ, Nordstrom RJ, Buatti J, Hylton N. Computational challenges and collaborative projects in the NCI quantitative imaging network. Tomography. 2016;2:242–249.
24. Alluri KC, Tahari AK, Wahl RL, Koch W, Chung CH, Subramaniam RM. Prognostic value of FDG PET metabolic tumor volume in human papillomavirus-positive stage III and IV oropharyngeal squamous cell carcinoma. Am J Roentgenol. 2014;203:897–903.
25. Sridhar P, Mercier G, Tan J, Truong MT, Daly B, Subramaniam RM. FDG PET metabolic tumor volume segmentation and pathologic volume of primary human solid tumors. AJR Am J Roentgenol. 2014;202:1114–1119.
26. Dibble EH, Alvarez AC, Truong MT, Mercier G, Cook EF, Subramaniam RM. 18F-FDG metabolic tumor volume and total glycolytic activity of oral cavity and oropharyngeal squamous cell cancer: adding value to clinical staging. J Nucl Med. 2012;53:709–715.
27. Foster B, Bagci U, Mansoor A, Xu Z, Mollura DJ. A review on segmentation of positron emission tomography images. Comput Biol Med. 2014;50:76–96.
28. Traverso A, Wee L, Dekker A, Gillies R. Repeatability and reproducibility of radiomic features: a systematic review. Int J Radiat Oncol Biol Phys. 2018;102:1143–1158.
29. Bailly C, Bodet-Milin C, Couespel S, Necib H, Kraeber-Bodéré F, Ansquer C, Carlier T. Revisiting the robustness of PET-based textural features in the context of multi-centric trials. PLoS ONE. 2016;11:e0159984.
30. Altazi BA, Zhang GG, Fernandez DC, Montejo ME, Hunt D, Werner J, Biagioli MC, Moros EG. Reproducibility of F18-FDG PET radiomic features for different cervical tumor segmentation methods, gray-level discretization, and reconstruction algorithms. J Appl Clin Med Phys. 2017;18:32–48.
31. Kalpathy-Cramer J, Mamomov A, Zhao B, Lu L, Cherezov D, Napel S, Echegaray S, Rubin D, McNitt-Gray M, Lo P, Sieren JC, Uthoff J, Dilger SK, Driscoll B, Yeung I, Hadjiiski L, Cha K, Balagurunathan Y, Gillies R, Goldgof D. Radiomics of lung nodules: a multi-institutional study of robustness and agreement of quantitative imaging features. Tomography. 2016;2:430–437.
32. Lu L, Liang Y, Schwartz LH, Zhao B. Reliability of radiomic features across multiple abdominal CT image acquisition settings: a pilot study using ACR CT phantom. Tomography. 2019;5:226–231.
33. Beichel RR, Smith BJ, Bauer C, Ulrich EJ, Ahmadvand P, Budzevich MM, Gillies RJ, Goldgof D, Grkovski M, Hamarneh G, Huang Q, Kinahan PE, Laymon CM, Mountz JM, Muzi JP, Muzi M, Nehmeh S, Oborski MJ, Tan Y, Zhao B, Sunderland JJ, Buatti JM. Multi-site quality and variability analysis of 3D FDG PET segmentations based on phantom and clinical image data. Med Phys. 2017;44:479–496.
34. Clark K, Vendt B, Smith K, Freymann J, Kirby J, Koppel P, Moore S, Phillips S, Maffitt D, Pringle M, Tarbox L, Prior F. The Cancer Imaging Archive (TCIA): maintaining and operating a public information repository. J Digit Imaging. 2013;26:1045–1057.
35. Beichel R, Ulrich EJ, Bauer C, Wahle A, Brown B, Chang T, Plichta KA, Smith BJ, Sunderland JJ, Braun T, Fedorov A, Clunie D, Onken M, Magnotta V, Menda Y, Riesmeier J, Pieper S, Kikinis R, Graham MM, Casavant TL, Sonka M, Buatti JM. Data From QIN-HEADNECK. The Cancer Imaging Archive; 2015. Available from: https://doi.org/10.7937/K9/TCIA.2015.K0F5CGLI.
36. Fedorov A, Beichel R, Kalpathy-Cramer J, Finet J, Fillion-Robin JC, Pujol S, Bauer C, Jennings D, Fennessy F, Sonka M, Buatti J, Aylward S, Miller JV, Pieper S, Kikinis R. 3D slicer as an image computing platform for the quantitative imaging network. Magn Reson Imaging. 2012;30:1323–1341.
37. Ulrich EJ, van To IM, Bauer C, Fedorov A, Beichel R. *PET Tumor Segmentation Extension Documentation 2017*. 2017. Available from: https://www.slicer.org/wiki/Documentation/Nightly/Extensions/PETTumorSegmentation.
38. Harrell FE, Jr., Califf RM, Pryor DB, Lee KL, Rosati RA. Evaluating the yield of medical tests. JAMA. 1982;247:2543–2546.
39. Obuchowski NA. An ROC-type measure of diagnostic accuracy when the gold standard is continuous-scale. Stat Med. 2006;25:481–493.
40. Schwartz LH, Litière S, de Vries E, Ford R, Gwyther S, Mandrekar S, Shankar L, Bogaerts J, Chen A, Dancey J, Hayes W, Hodi FS, Hoekstra OS, Huang EP, Lin N, Liu Y, Therasse P, Wolchok JD, Seymour L. RECIST 1.1-update and clarification: from the RECIST committee. Eur J Cancer. 2016;62:132–137.
41. Therasse P, Arbuck SG, Eisenhauer EA, Wanders J, Kaplan RS, Rubinstein L, Verweij J, Van Glabbeke M, van Oosterom AT, Christian MC, Gwyther SG. New guidelines to evaluate the response to treatment in solid tumors. European Organization for Research and Treatment of Cancer, National Cancer Institute of the United States, National Cancer Institute of Canada. J Natl Cancer Inst. 2000;92:205–216.
42. Echegaray S, Gevaert O, Shah R, Kamaya A, Louie J, Kothary N, Napel S. Core samples for radiomics features that are insensitive to tumor segmentation: method and pilot study using CT images of hepatocellular carcinoma. J Med Imaging. 2015;2:041011.
43. Echegaray S, Nair V, Kadoch M, Leung A, Rubin D, Gevaert O, Napel S. A rapid segmentation-insensitive "digital biopsy" method for radiomic feature extraction: method and pilot study using CT images of non-small cell lung cancer. Tomography. 2016;2:283–294.
44. van Timmeren JE, Leijenaar RTH, van Elmpt W, Wang J, Zhang Z, Dekker A, Lambin P. Test-retest data for radiomics feature stability analysis: generalizable or study-specific? Tomography. 2016;2:361–365.
45. Wahl RL, Jacene H, Kasamon Y, Lodge MA. From RECIST to PERCIST: evolving considerations for PET response criteria in solid tumors. J Nucl Med. 2009;50:122s–150s.
46. Mertens J, Dobbeleir A, Ham H, D'Asseler Y, Goethals I, Van de Wiele C. Standardized added metabolic activity (SAM): a partial volume independent marker of total lesion glycolysis in liver metastases. Eur J Nucl Med Mol Imaging. 2012;39:1441–1448.