NATIONAL & KAPODISTRIAN UNIVERSITY OF ATHENS

SCHOOL OF SCIENCES
DEPARTMENT OF INFORMATICS AND TELECOMMUNICATIONS

INTERDEPARTMENTAL PROGRAM OF POSTGRADUATE STUDIES IN DATA SCIENCE AND INFORMATION TECHNOLOGIES
SPECIALIZATION: BIOINFORMATICS – BIOMEDICAL DATA SCIENCE

MASTER THESIS

Predicting Head and Neck Cancer Patients' Survival Using Computed Tomography-Derived Skeletal Muscle Related Data

Paris T. Moumoulidis

Supervisor: Vassilis Katsouros, Research Director of Institute for Language and Speech Processing (ILSP), ATHENA Research and Innovation Center

ATHENS

DECEMBER 2023

ΕΘΝΙΚΟ & ΚΑΠΟΔΙΣΤΡΙΑΚΟ ΠΑΝΕΠΙΣΤΗΜΙΟ ΑΘΗΝΩΝ

ΣΧΟΛΗ ΘΕΤΙΚΩΝ ΕΠΙΣΤΗΜΩΝ
ΤΜΗΜΑ ΠΛΗΡΟΦΟΡΙΚΗΣ ΚΑΙ ΤΗΛΕΠΙΚΟΙΝΩΝΙΩΝ

ΔΙΑΤΜΗΜΑΤΙΚΟ ΠΡΟΓΡΑΜΜΑ ΜΕΤΑΠΤΥΧΙΑΚΩΝ ΣΠΟΥΔΩΝ ΣΤΗΝ ΕΠΙΣΤΗΜΗ ΔΕΔΟΜΕΝΩΝ ΚΑΙ ΤΕΧΝΟΛΟΓΙΕΣ ΠΛΗΡΟΦΟΡΙΑΣ
ΕΙΔΙΚΕΥΣΗ: ΒΙΟΠΛΗΡΟΦΟΡΙΚΗ – ΕΠΙΣΤΗΜΗ ΒΙΟΪΑΤΡΙΚΩΝ ΔΕΔΟΜΕΝΩΝ

ΔΙΠΛΩΜΑΤΙΚΗ ΕΡΓΑΣΙΑ

Προβλέποντας την Επιβίωση σε Ασθενείς με Καρκίνο Κεφαλής και Τραχήλου Αξιοποιώντας Δεδομένα Σχετικά με τους Σκελετικούς Μύες από την Αξονική Τομογραφία

Πάρις Θ. Μουμουλίδης

Επιβλέπων: Βασίλειος Κατσούρος, Διευθυντής Ερευνών Ινστιτούτου Επεξεργασίας του Λόγου, Ερευνητικό Κέντρο ”Αθηνά”

ΑΘΗΝΑ

ΔΕΚΕΜΒΡΙΟΣ 2023

MASTER THESIS

Predicting Head and Neck Cancer Patients' Survival Using Computed Tomography-Derived Skeletal Muscle Related Data

Paris T. Moumoulidis

S.N.: DS2200012

SUPERVISOR:

Vassilis Katsouros, Research Director of Institute for Language and Speech Processing (ILSP), ATHENA Research and Innovation Center

EXAMINATION COMMITTEE:

Vassilis Katsouros, Research Director of Institute for Language and Speech Processing (ILSP), ATHENA Research and Innovation Center

Vassilis Papavasileiou, Research Associate of Institute for Language and Speech Processing (ILSP), ATHENA Research and Innovation Center

Efthymios Kyrodimos, Associate Professor, Medical School, National and Kapodistrian University of Athens

ΔΙΠΛΩΜΑΤΙΚΗ ΕΡΓΑΣΙΑ

Προβλέποντας την Επιβίωση σε Ασθενείς με Καρκίνο Κεφαλής και Τραχήλου Αξιοποιώντας Δεδομένα Σχετικά με τους Σκελετικούς Μύες από την Αξονική Τομογραφία

Πάρις Θ. Μουμουλίδης

Α.Μ.: DS2200012

ΕΠΙΒΛΕΠΩΝ:

Βασίλης Κατσούρος, Διευθυντής Ερευνών Ινστιτούτου Επεξεργασίας του Λόγου, Ερευνητικό Κέντρο ”Αθηνά”

ΕΞΕΤΑΣΤΙΚΗ ΕΠΙΤΡΟΠΗ:

Βασίλης Κατσούρος, Διευθυντής Ερευνών Ινστιτούτου Επεξεργασίας του Λόγου, Ερευνητικό Κέντρο ”Αθηνά”

Βασίλης Παπαβασιλείου, Επιστημονικός συνεργάτης Ινστιτούτου Επεξεργασίας του Λόγου, Ερευνητικό Κέντρο ”Αθηνά”

Κυροδήμος Ευθύμιος, Αναπληρωτής Καθηγητής, Ιατρική Σχολή, Εθνικό και Καποδιστριακό Πανεπιστήμιο Αθηνών

ABSTRACT

Objective: The purpose of the project is to propose a machine learning based classification model, able to identify patients in high risk for decreased overall survival based only on CT-derived muscle related data, in patients with stage IV HNSCCs. As part of the project, an automated paravertebral muscle area (with and without intermuscular and intramuscular adipose tissue) segmentation method will be developed and proposed. Our aim will not be to achieve near perfect classification results (something utopic due to the complex medical background of the problem addressed), but to identify possibly high-risk group of patients that may be benefited from targeted nutritional and other interventions. Therefore, we are aiming to develop an automated screening method that will be based on CT-derived muscle related data. **Material and Methods:** A PET-CT collection, with 298 patients with histologically proven head-and-neck cancer, was retrieved from the cancer imaging archive and was used for the purposes of this pilot study. We included only patients with Stage IV cancer, with known site of the primary tumour and with a minimum follow-up period of 5 years. These inclusion criteria resulted in 74 patients. Further sub-cohorts (with 47 and 51 patients) were created with the application of extra exclusion criteria in the group of patients with oropharyngeal carcinomas. Premature death was defined as death when the survival probability was higher than 75% in the separate, for each primary site, survival curves. Unsupervised machine learning methods were also used to address the separability of our data and to test different feature selection strategies. Classification results after training on both manually and automatically segmented muscle areas were evaluated. Best performing classifiers were tested on a validation set consisted of the three images per patient that had not been used for training. Validation results were tested in terms of classifiers' ability to separate survival curves of the low-risk and the high-risk group of patients statistically significantly. Survival analysis was performed using Kaplan-Meier survival curves. **Results:** In unsupervised learning we observed that when excluding patients with OPSCC without premature death, there seemed to be an inherent 3-cluster tendency in our dataset (one cluster with overrepresentation of low-risk patients and two clusters with overrepresentation of high-risk patients). Our classification results were very encouraging, as we managed to train classifiers that served well the screening purposes of the problem addressed, by achieving high recall while maintaining an acceptable F1-score. The best results in the validation set were obtained in the cohort with 47 patients and when classification models were trained with 7 principal components and with a test ratio of 0.3. A soft voting ensemble model achieved to showcase a trend for difference in survival curves between the two risk groups (p-value < 0.1) in 80% of the 40 different train-test splits of the dataset, and to separate statistically significantly the two curves in 65% of the splits. **Conclusion:** The proposed automatic method for segmentation, radiomic feature extraction and subsequent patient risk stratification, based on CT-derived skeletal muscle related data, constitutes a promising automatic screening method. The fact that results were evaluated on 40 different train-test splits of the dataset and that proposed risk stratification was tested on a validation set using the same risk cut-off points and not always the optimal ones, along with the consistency regarding various classifiers' performance pave the way for potential generalization. However, more data are needed to establish risk stratification based on CT-derived skeletal muscle related data as a clinically useful biomarker.

SUBJECT AREA: Radiomics-based machine learning

KEYWORDS: Radiomics, head and neck cancer, automatic segmentation, risk stratification, machine learning

ΠΕΡΙΛΗΨΗ

Σκοπός: Η μελέτη στοχεύει στο να προτείνει ένα μοντέλο ταξινόμησης μηχανικής μάθησης ικανό να αναγνωρίζει ασθενείς υψηλού ρίσκου για μειωμένη συνολική επιβίωση, βασιζόμενο μόνο σε δεδομένα σχετικά με τους σκελετικούς μύες από την αξονική τομογραφία, σε ασθενείς με σταδίου 4 καρκίνο της κεφαλής και του τραχήλου. Ως μέρος της μελέτης θα αναπτυχθεί και θα προταθεί μία μέθοδος αυτόματης κατάτμησης της περιοχής ενδιαφέροντος στην αξονική τομογραφία των παρασπονδυλικών μυών (με και χωρίς το περιμυϊκό και ενδομυϊκό λιπώδη ιστό). Στοχεύουμε στο να αναπτύξουμε μια μέθοδο διαλογής των ασθενών υψηλού κινδύνου που θα μπορούσαν να ωφεληθούν από διατροφικές ή άλλες παρεμβάσεις, βασιζόμενη σε δεδομένα σχετικά με τους σκελετικούς μύες από την αξονική τομογραφία, και όχι να πετύχουμε κοντά στο τέλειο αποτελέσματα ταξινόμησης, κάτι που ούτως ή άλλως είναι ουτοπικό εξαιτίας του πολύπλοκου ιατρικού υποβάθρου του προβλήματος που απευθύνουμε. **Υλικό και Μέθοδος:** Αποκτήσαμε πρόσβαση σε μία συλλογή PET-CT του αρχείου απεικονίσεων καρκίνου της TCIA που περιλάμβανε 298 ασθενείς με ιστολογικώς αποδεδειγμένο καρκίνο της κεφαλής και του τραχήλου. Στη μελέτη συμπεριλάβαμε μόνο ασθενείς σταδίου 4 , με γνωστή πρωτοπαθή εστία και με ελάχιστη περίοδο παρακολούθησης τα 5 έτη, καταλήγοντας έτσι σε 74 ασθενείς. Με την εφαρμογή περαιτέρω κριτηρίων αποκλεισμού στη κατηγορία των ασθενών με καρκίνο του στοματοφάρυγγα δημιουργήθηκαν μικρότερες κοορτές των 47 και 51 ασθενών. Ως πρόωρος θάνατος ορίστηκε ξεχωριστά για ασθενείς με διαφορετική πρωτοπαθή εστία , ο θάνατος όταν η πιθανότητα επιβίωσης στις καμπύλες επιβίωσης ήταν μεγαλύτερη του 75%. Χρησιμοποιήσαμε ακόμη μεθόδους μη επιβλεπόμενης μάθησης προκειμένου να δούμε την έμφυτη τάση των δεδομένων μας για διαχωρισμό σε ομάδες, καθώς και για να τεστάρουμε διαφορετικές στρατηγικές επιλογής χαρακτηριστικών. Τα αποτελέσματα ταξινόμησης μετά την εκπαίδευση των μοντέλων αξιολογήθηκαν τόσο στις εικόνες που είχε γίνει χειροκίνητα η κατάτμηση των περιοχών ενδιαφέροντος των μυών όσο και στις εικόνες με αυτόματη κατάτμηση. Οι ταξινομητές με τα καλύτερα αποτελέσματα αξιολογήθηκαν σχετικά με την ικανότητά τους να κατηγοριοποιούν τους ασθενείς σε υψηλού και χαμηλού ρίσκου με τρόπο ώστε να χωρίζουν σε βαθμό στατιστικά σημαντικό οι καμπύλες επιβίωσης μεταξύ των δύο ομάδων ρίσκου των ασθενών. Η ανάλυση επιβίωσης έγινε χρησιμοποιώντας τις κατά Kaplan-Meier καμπύλες επιβίωσης. **Αποτελέσματα:** Χρησιμοποιώντας μεθόδους μη επιβλεπόμενης μάθησης παρατηρήσαμε ότι αποκλείοντας ασθενείς με καρκίνο του στοματοφάρυγγα χωρίς πρόωρο θάνατο, υπήρχε μια έμφυτη τάση για σχηματισμό 3 ομάδων (1 με σαφή κυριαρχία των ασθενών χαμηλού ρίσκου και 2 όπου κυριαρχούσαν οι ασθενείς υψηλού ρίσκου). Τα αποτελέσματα επιβλεπόμενης μάθησης ήταν επίσης πολύ ενθαρρυντικά, επιτυγχάνοντας εξαιρετική ευαισθησία διατηρώντας αποδεκτά F1-score. Τα καλύτερα αποτελέσματα επιτεύχθηκαν στην κοορτή με 47 ασθενείς, όταν η εκπαίδευση έγινε χρησιμοποιώντας 7 κύριες συνιστώσες, αφήνοντας για τεστ 30% των δεδομένων, με το καλύτερο μοντέλο να καταφέρνει να αναδείξει τάση διαφοροποίησης των καμπυλών επιβίωσης των δύο ομάδων κινδύνου στο 80% των 40 διαφορετικών διαχωρισμών για εκπαίδευση-αξιολόγηση των δεδομένων. **Συμπεράσματα:** Η προτεινόμενη μέθοδος αυτόματης κατάτμησης της περιοχής ενδιαφέροντος, εξαγωγής ραδιομικών χαρακτηριστικών και διαστρωμάτωσης κινδύνου των ασθενών είναι πολλά υποσχόμενη, με δυναμικό γενίκευσης, ωστόσο απαιτούνται περισσότερα δεδομένα πριν προταθεί ως χρήσιμος στην κλινική πρακτική βιοδείκτης.

ΘΕΜΑΤΙΚΗ ΠΕΡΙΟΧΗ: Μηχανική μάθηση βασισμένη σε ραδιομική ανάλυση

ΛΕΞΕΙΣ ΚΛΕΙΔΙΑ: Ραδιομική ανάλυση, καρκίνος κεφαλής και τραχήλου, αυτόματη κατάτμηση, διαστρωμάτωση κινδύνου, μηχανική μάθηση

Στη Χλόη, την Κωνσταντίνα και τον Γιώργο που ήταν η καλύτερη ομάδα που θα μπορούσα να φανταστώ και από τους οποίους έμαθα τόσα, καθώς και στον φίλο μου Γιάννη που με προέτρεψε να ξεκινήσουμε αυτό το μεταπτυχιακό

CONTENTS

LIST OF FIGURES

LIST OF TABLES

1. INTRODUCTION

1.A Purpose of the study

The majority of patients with head and neck squamous cell carcinomas (HNSCC) present with locoregionally advanced stages (III, IV). In these group of patients, notably in stage IV patients, multimodal interventions are required, as well as personalized treatment options, in order to minimize the side effects of the treatment applied, while achieving increased overall survival (OS) and superior quality of life (QoL) [1], [2].

The nutritional status of the patient (as expressed by sarcopenia, myosteatosis, frailty and others) together with the presence of human papilloma virus (HPV) infection hold great prognostic significance in those patients (see chapter 1.B). Therefore, it is important to identify patients in which treatment benefits outweigh the risk of any adverse outcome, as well as patients who are in high risk for decreased overall survival. However, it can be not only highly time consuming to assess frailty and sarcopenia, but also sometimes impossible (especially in retrospective studies) due to the lack of critical clinical information required. Moreover, there is no consensus regarding cut off values for sarcopenia in literature with studies done in different parts of the world citing different cut off values.

The aforementioned limitations lead us to investigate the feasibility of extracting data related to sarcopenia, myosteatosis and frailty from the patient's CT-scan, which is routinely performed, and subsequently available in all patients. Based on the literature (see 1B.4) we decided to extract these radiomic information from the paravertebral muscle area at the level of the third cervical vertebra (C3). Manual paravertebral muscle segmentation remains another time-consuming process, therefore there is need for automated segmentation methods.

The purpose of the project is to propose a machine learning based classification model, able to identify patients in high risk for decreased overall survival based only on CT-derived muscle related data, in patients with stage IV HNSCCs. As part of the project, an automated paravertebral muscle area (with and without intermuscular and intramuscular adipose tissue) segmentation method will be developed and proposed. Classification results after training on both manually and automatically segmented muscle areas shall be evaluated. Unsupervised machine learning methods will be also used in order to address the separability of our data and in order to test different feature selection strategies.

Given the fact that survival is affected by various other, both known and yet unknown factors, we will focus on the occurrence of premature death, while taking into account the site of the primary tumour. Especially, we shall treat with great precaution data from oropharyngeal cancer (OPC) patients, as in this specific group, survival is highly affected by the HPV status, which is not always available. Our aim will not be to achieve near perfect classification results (something utopic due to the complex medical background of the problem addressed), but to identify possibly high-risk group of patients that may be benefited from targeted nutritional and other interventions. Therefore, we are aiming to develop an automated screening method that will be based on CT-derived muscle related data. Moreover the muscle related risk stratification of the patients, may serve as an extra feature in more complicated prognostic models that will also include well established prognostic factors.

1.B Definitions, medical background, and relative research interest in Otorhinolaryngology

1.B.1 Head-and-neck squamous cell carcinoma

Head and neck squamous cell carcinomas (HNSCCs) develop from the mucosal epithelium of the oral cavity, nasopharynx, oropharynx, hypopharynx and larynx [3]. Despite the similar cell origin, tumours of the nasopharynx constitute a separate epithelial malignancy entity with distinct geographical distribution and different, compared to other epithelial head and neck tumours, pathogenesis progression and aetiology (including EBV infection, host genetics, and environmental factors) [4]. Due to the aforementioned differences nasopharyngeal carcinomas are studied separately. Sinonasal squamous-cell carcinomas are studied separately as well, as they constitute a quite complex tumour type for with numerous histologic variants and unusual morphologic features, with their aetiology, epidemiology, clinical features, and genetic profiles being quite distinct from those of the main head and neck cancer localizations [5]. The main anatomical sites of HNSCC development are shown in Figure 1 [3].

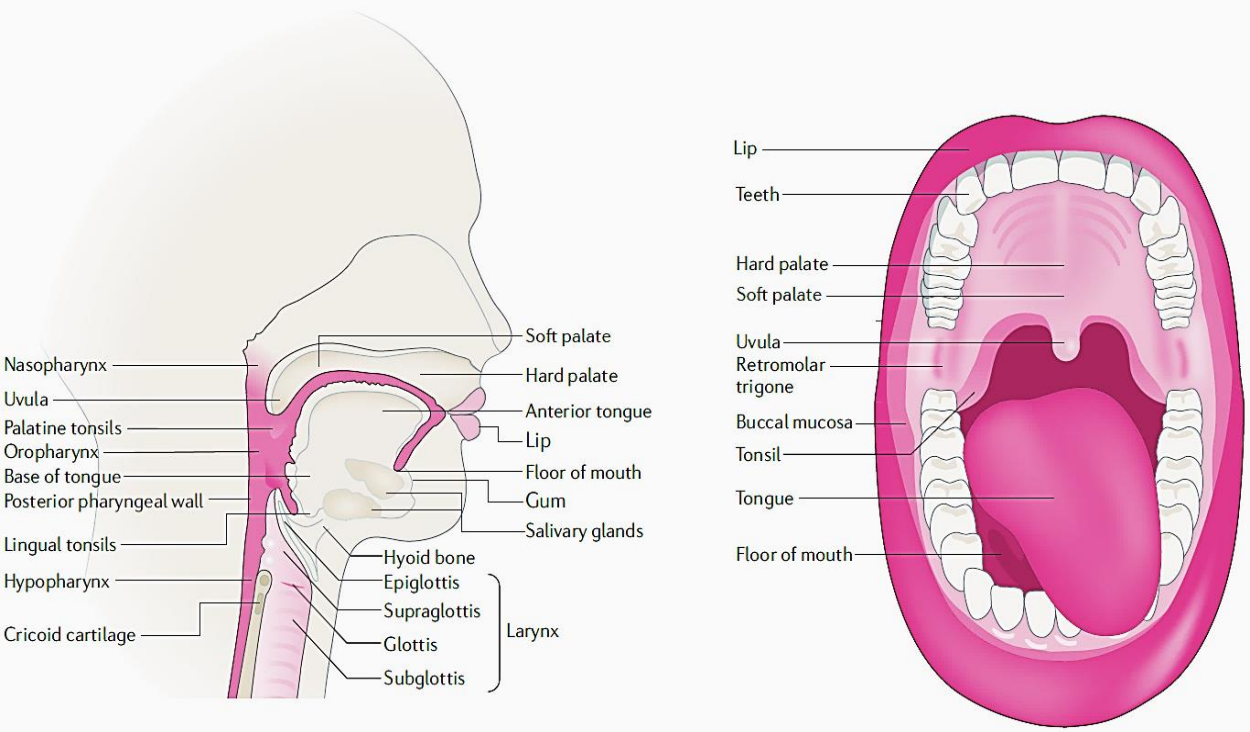


Figure 1: Head and neck squamous cell carcinoma (HNSCC) arises from the mucosal epithelium of the oral cavity (lips, buccal mucosa, hard palate, anterior tongue, floor of mouth and retromolar trigone), nasopharynx, oropharynx (palatine tonsils, lingual tonsils, base of tongue, soft palate, uvula and posterior pharyngeal wall), hypopharynx (the bottom part of the throat, extending from the hyoid bone to the cricoid cartilage) and larynx

It has been almost 20 years now that trends in head and neck cancer (HNC) have started changing. It has been reported [6] that oral cavity cancer incidence rates increased in many countries with peaking tobacco epidemics and on the other hand declined in countries where tobacco use peaked some time ago. Moreover, rates of oropharyngeal cancer increased in a number of countries where tobacco use has declined [6], and the incidence of human papilloma virus (HPV) associated oropharyngeal cancer is increasing in developed countries [7], induced mainly by HPV type 16, affecting predominantly younger people in North America and northern Europe, reflecting a latency of 10 to 30

years after oral-sex exposure [8]. Despite the distinct patterns of geographic variation in HPV-related oropharyngeal cancer, with higher prevalence in Western Europe, there are still limited recent data available for Eastern Europe, Asia or Africa [9].

Hence, HNSCC can be classified into two distinct types, HPV-positive and HPV-negative, with distinct mutational landscape, response to clinical treatment, and survival outcomes [7]. HPV-associated HNSCCs arise primarily from the palatine and lingual tonsils of the oropharynx, whereas tobacco-associated HNSCCs arise primarily in the oral cavity, hypopharynx and larynx [3]. The main characteristics of these two distinct types of HNSCC are summarized in Table 1 [7], [10] .

Table 1: HPV-negative compared with HPV-positive head and neck cancer

Parameter	**HPV (-)**	**HPV (+)**
Male:Female	3:1	8:1
Age	> 60	40-60
Race	White > black; worse prognosis in blacks	White > black
Socioeconomic status	Low-middle	Higher
Smoking	>90% have smoking history; risk increases with increasing tobacco use	50%–65% have smoking history
Alcohol consumption	Synergistic with tobacco in increasing risk	Not a significant risk factor
Sexual history	Not a significant risk factor	Number of oral sex partners is a strong risk factor
Primary tumour site	Larynx and oral cavity most common	Oropharynx, specifically lymphoid tissue of tonsils and tongue base
Presentation	Varies	Enlarged cervical lymph nodes common; also oropharyngeal pain, dysphagia, referred otalgia
Incidence trends	Decreasing	Increasing
Tumour (T) stage	More advanced T stage	Early T stage
Nodal (N) stage	Early N stage	Advanced N stage
Second primary rate (%)	4.6	11
Prognosis	All sites: 5-year survival 65%, 5-year recurrence 50% Oropharynx: 5-year survival 20%–25%, 5-year recurrence 50%	Oropharynx: 5-year survival 60%–90%, 5-year recurrence 10%–15%

The need for continued awareness in reducing HNSCC traditional risks factors, such as cigarette use, remains, while emerging risk factors like HPV infection, require novel staging systems and greater resources to be poured into, in order to decrease the incidence of HNSCC worldwide [11]. Despite the significant drop in the incidence of smoking-related HNSCC, efforts to decrease cigarette usage should continue, and newly emerged potential risk factors, such as the E-cigarettes, should be tackled, given that they are ineffective in helping head and neck cancer patients achieve smoking cessation [12], E-cigarettes' exact role in HNSCC development has not been clarified [13], and their young users are more likely to use conventional cigarettes in the future [14].

1.B.2 Definitions

1.B.2.1 Sarcopenia

In 2018, the European Working Group on Sarcopenia in Older People met for the second time (EWGSOP2) in order to update the original definition of sarcopenia, in a way that would reflect all the scientific and clinical evidence built since their first published definition back in 2010. Aiming to increase consistency of research design, clinical diagnoses and ultimately, care for people with sarcopenia, they proposed the following operational definition of sarcopenia [15] : Sarcopenia is probable when low muscle strength is detected. A sarcopenia diagnosis is confirmed by the presence of low muscle quantity or quality. When low muscle strength, low muscle quantity/quality and low physical performance are all detected, sarcopenia is considered severe. As a way of applying this definition in practice, EWGSOP2 reviewed tests and tools used for assessing muscle properties and performance, and presented the EWGSOP2 sarcopenia cut-off points, as seen in Table 2 [15], and an updated algorithm for sarcopenia case-finding, diagnosis and severity determination as seen in Figure 2 [15].

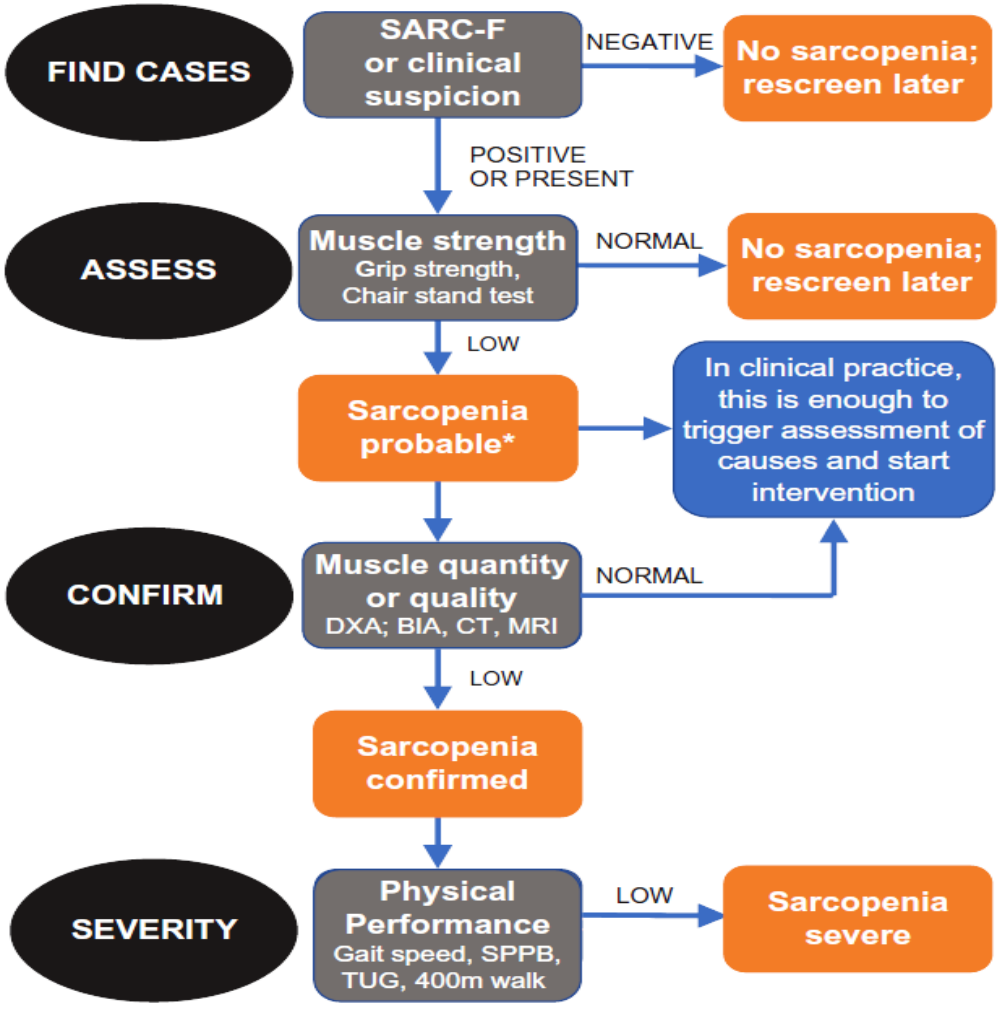


Figure 2: Sarcopenia: EWGSOP2 algorithm for case-finding, making a diagnosis and quantifying severity in practice. The steps of the pathway are represented as Find-Assess-Confirm-Severity or F-A-C-S. *Consider other reasons for low muscle strength (e.g. depression, stroke, balance disorders, peripheral vascular disorders).

Table 2: EWGSOP2 sarcopenia cut-off points

Test	Cut-off points for men	Cut-off points for women	Reference
EWGSOP2 sarcopenia cut-off points for low strength by chair stand and grip strength			
Grip strength	<27 kg	<16kg	[16]
Chair stand	>15 s for five rises		[17]
EWGSOP2 sarcopenia cut-off points for low muscle quantity			
Appendicular skeletal muscle mass (ASM)	<20kg	<15kg	[18]
ASM/height^2	<7.0kg/m^2	<5.5kg/m^2	[19]
EWGSOP2 sarcopenia cut-off points for low performance			
Gait speed	≤0.8 m/s		[20], [21]
SPPB	≤8 point score		[22], [23]
TUG	≥20 s		[24]
400 m walk test	Non-completion or ≥6 min for completion		[25]

1.B.2.2 Frailty

Frailty is a state of vulnerability to poor resolution of homoeostasis after a stressor event and is a consequence of cumulative decline in multiple body systems or functions (physical, cognitive, social, and psychological) during a lifetime, increasing susceptibility to poor health outcomes and remaining the most problematic expression of population ageing [26]. While the physical phenotype of frailty shows significant overlap with sarcopenia with low grip strength, slow gait speed and weight loss being involved in both, frailty and sarcopenia are still distinct— one a geriatric syndrome representing a much broader concept and the other a disease [15].

In 2013 a frailty consensus from 6 major international, European, and US societies was published [27], highlighting the following 4 key points regarding physical frailty:

1. Physical frailty is an important medical syndrome. Physical frailty was defined as "a medical syndrome with multiple causes and contributors that is characterized by diminished strength, endurance, and reduced physiologic function that increases an individual's vulnerability for developing increased dependency and/or death."

2. Physical frailty can potentially be prevented or treated with specific modalities, such as exercise, protein-calorie supplementation, vitamin D, and reduction of polypharmacy.

3. Simple, rapid screening tests have been developed and validated, such as the simple FRAIL scale, to allow physicians to objectively recognize frail persons.

4. For the purposes of optimally managing individuals with physical frailty, all persons older than 70 years and all individuals with significant weight loss (≥5%) due to chronic disease should be screened for frailty.

Frailty's diagnostic tools were built based on two main models: 1) the phenotypic model, which describes a relationship between a set of criteria that defines frailty and the effect on certain outcomes, and 2) the deficit accumulation model, which measures the number

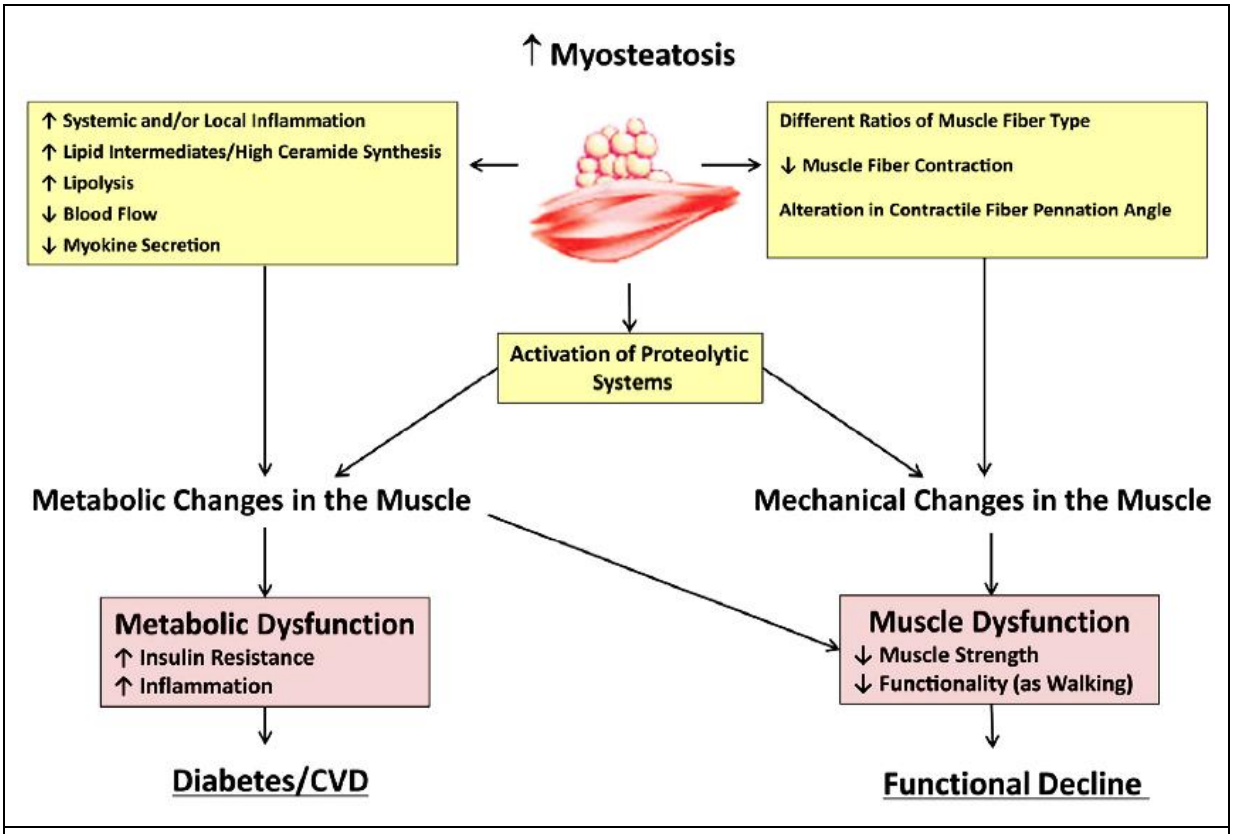
of deficits that an individual has accrued across a number of different domains, including comorbidities, the ability to manage activities of daily living, and physical signs [28]. Fried's phenotypic model [29] is a predominantly physical conceptualization and is based on evaluating unintentional weight loss (shrinking), grip strength (weakness), self-reported exhaustion (poor endurance and energy), slow walking speed (slowness), and low physical activity. On the other hand, the Frailty Index (FI) [30] consists of a 70-item scale derived from history and physical exam and is calculated as a ratio of the possible number of deficits (up to 70) to the number of actual deficits present in the individual.

Nowadays, comprehensive geriatric assessment (CGA) that evaluates physical, psychological, functional, and social capabilities, and limitations of geriatric patients is the accepted gold standard for caring for frail older people in hospital, with a recently published umbrella review [31] highlighting that there is a degree of consistency in definition, essential content, key target group and outcomes of CGA. However, such assessments are time-consuming, leading many cancer specialists to seek a shorter screening tool that can separate fit older adults with cancer, who can receive standard cancer treatment, from vulnerable patients, who should subsequently receive a full assessment to guide tailoring of their treatment regimens [32]. One such tool is the Geriatrics 8 (G8) screening tool, which was developed specifically for older adult patients with cancer. G8 consists of eight items which cover multiple geriatric domains, including nutritional status, physical capacity, mood, and polypharmacy. G8 scores range from zero to seventeen, with scores ≤ fourteen representing potential frailty [33].

1.B.2.3 Myosteatosis

Myosteatosis occurs as a result of fatty infiltration of skeletal muscle tissue. An interdisciplinary workshop convened by the National Institute on Aging Division of Geriatrics and Clinical Gerontology on September 2018, discussed myosteatosis in the context of skeletal muscle function deficit (SMFD)[34].Traditionally the term myosteatosis has been used to describe multiple different adipose depots found in skeletal muscle including: (a) intermuscular adipose tissue, the extracellular adipose tissue found beneath the fascia and in-between muscle groups; (b) intramuscular adipose tissue, the extracellular adipose tissue found within an individual muscle; and (c) intramyocellular lipids. Intermuscular, intramuscular, and intramyocellular fat all provide a slightly different measure of myosteatosis and may represent different risk factors to metabolic and muscle health particularly in older adults. In myosteatosis ectopic fat depot increases with aging and is recognized to negatively correlate with muscle mass, strength, and mobility and disrupt metabolism (insulin resistance, diabetes). Figure 3 [35] demonstrates pathophysiology changes in myosteatosis.

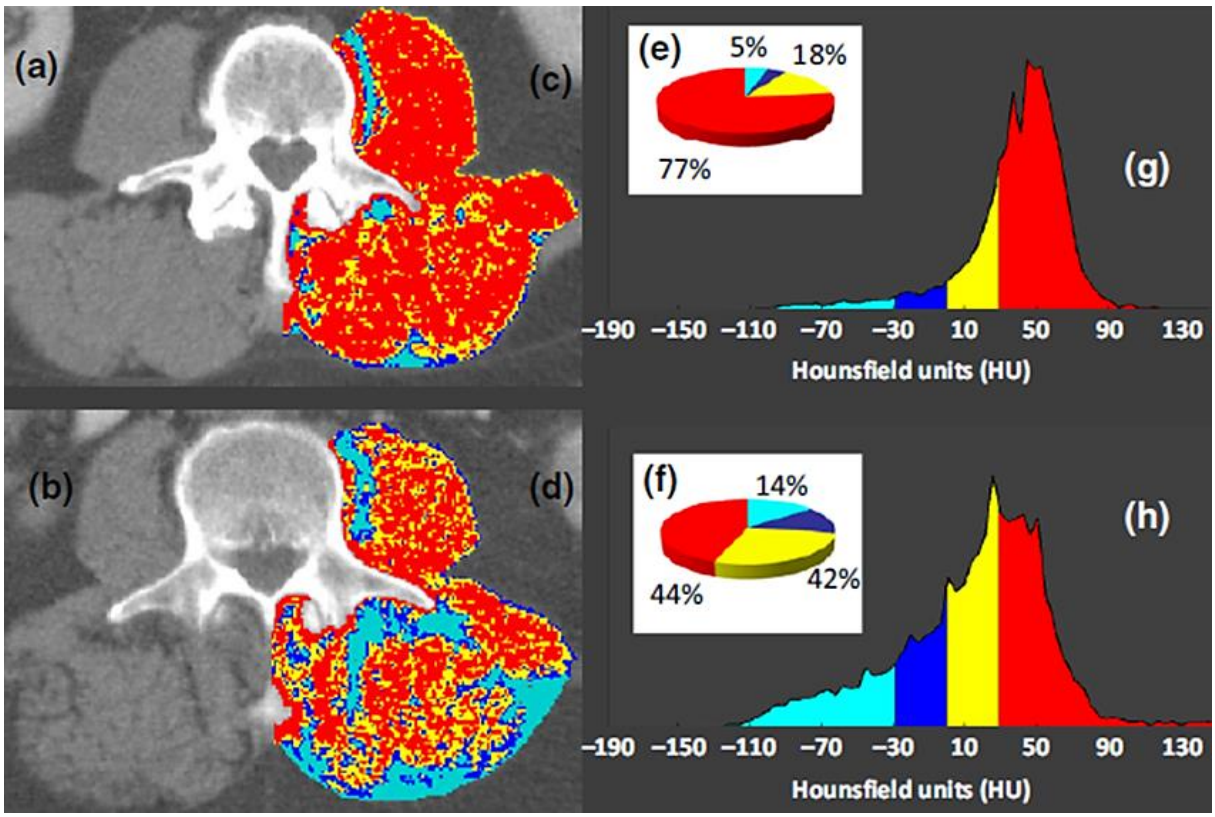
Although myosteatosis is not synonymous with sarcopenia (loss of muscle mass and function), it does appear to be independent of muscle mass and perhaps act synergistically. Studying myosteatosis role as a newly defined independent risk factor should be expanded. Opportunistic opportunities like cancer populations, shoulder injury patients, bariatric surgery patients, and conditions that may accelerate myosteatosis would considerably expand our knowledge and open an array of research prospects in the field.



Increased myosteatosis may lead to metabolic and mechanical changes in the muscle through a variety of mechanisms. Changes in muscle cell metabolism can lead to increased insulin resistance and inflammation, aiding in the development of diabetes, and cardiovascular diseases. Alterations in muscle architecture can also lead to muscular dysfunction and functional decline. Both processes may be increased through activation of proteolytic systems, which may also result from increased myosteatosis.

Figure 3: Potential mechanisms underlying the effects of myosteatosis

In 2017, the need for standardized assessment of myosteatosis was discussed in a symposium[35]. Imaging methods that can easily and rapidly assess muscle composition in multiple clinical settings and with minimal patient burden, were discussed as well. Magnetic resonance imaging (MRI) is considered an excellent non-invasive technique for measuring myosteatosis, providing high-quality images, yet the cost is high, and traditional MRI does not typically allow quantification of the fat content of the muscle. Computed tomography (CT) on the other hand, has been the most utilized as a research tool to investigate myosteatosis. The CT analysis of myosteatosis is based on a Hounsfield unit (HU), which is a measure associated with the way rays pass through water. Water has a density of zero, higher measurements are denser (i.e., bone), and lower measurements are less dense (i.e., fat). The lower the density, the lower the Hounsfield units and the higher the degree of myosteatosis. Any given skeletal muscle displays radiation attenuation between -190 and +150 Hounsfield units (HU), with a prominent peak near +50 HU. When muscle cross-sectional area and attenuation are reported, the most common practice is to use predefined HU ranges to demarcate intermuscular adipose tissue (usually -190 to -30 HU) and muscle tissue (usually -29 HU to 150 HU) [36]. Figure 4 [36] demonstrates radiation attenuation map of paraspinal muscles with and without myosteatosis.



(a, c, e, g): Subject 1 has hardly visible intermuscular fat (4.6% of total tissue area), 77.2% of the total muscle cross-sectional area falls into the normal attenuation range for muscle and the mean overall radiation attenuation is 42.3 HU.

(b, d, f, h): Subject 2 exhibits extensive regions (14.1%) of intermuscular fat infiltration, low overall mean attenuation (20.4 HU) and less than half (44.4%) of the total tissue cross-sectional area falls within the normal range of muscle radiation attenuation values.

a,b: CT images of paraspinal muscles ;

c,d: annotated CT images;

e,f: pie charts;

g,h: histograms of radiation attenuation showing the percentages of total tissue cross-sectional area within the following attenuation ranges:

adipose tissue [light blue, -190 to -30 HU],

normal attenuation muscle [red, +30 to +150 HU],

abnormal (reduced) attenuation muscle in two ranges [dark blue, -29 to 0 HU; yellow, +1 to +29 HU]

Figure 4: Radiation attenuation map of paraspinal muscles with and without myosteatosis

1.B.2.4 Radiomics

The word omics refers to a field of study in biological sciences that ends with -omics, such as genomics, transcriptomics, proteomics, or metabolomics. The ending -ome is used to address the objects of study of such fields, such as the genome, proteome, transcriptome, or metabolome, respectively. In medicine all these "omics" concepts have resulted in an incremental growth of medical big data. In order to extract the desired information from all these emerging data, different techniques from artificial intelligence (AI), mainly

machine learning and deep learning algorithms, are increasingly being applied in the medical sector. Over the past decade, medical imaging analysis has grown exponentially [37] leading to the vigorous development of another "-omic" concept, called "radiomics". Radiomics is a quantitative approach to medical imaging, aiming at enhancing the existing data available to clinicians by means of advanced mathematical analysis [38]. Through mathematical extraction of the spatial distribution of signal intensities and pixel interrelationships, radiomics quantifies textural information, overcoming the subjective nature of image interpretation, and extracts quantitative features.

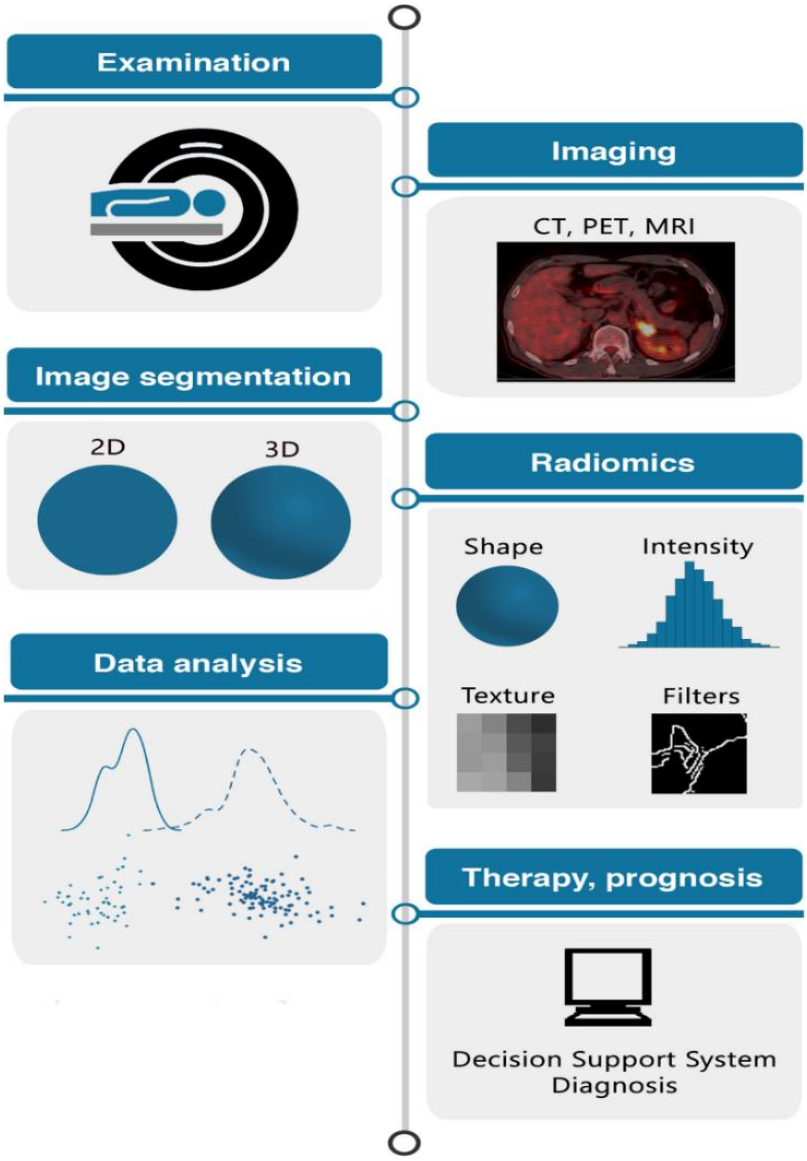


Figure 5: The radiomic workflow

Although radiomics can be applied to various conditions, it is most well developed in oncology [37], [39]. A very high number of features can be extracted from various imaging modalities, including CT, MRI and positron emission tomography (PET), alone or combined, contributing to better tumour and environment characterization, early detection of relapse after radical treatment and development of a patient's phenotype that could lead to personalized treatment. [40]. Furthermore, radiomic data can be combined with other relevant data, such as medical notes from electronic-health records, pathology, biology, or genomics, in an attempt to develop models that could improve diagnostic, prognostic, and predictive accuracy, facilitating better clinical decision making [37], [41].

Figure 5 [38] shows a schematic illustration of the radiomics workflow and Table 3 [40], [42] provides a brief overview of the major radiomic features.

Table 3: Overview of major radiomic feature

	Concept and aim	***Summary of statistics and characteristics***	***Clinical application in oncology***
Structural features	Basic descriptors according to physical characteristics (e.g., shape, volume, size).	Volume of region of interest (ROI), max axial length, max 3D diameter, surface area, surface-to-volume ratio, sphericity, compactness, spherical disproportion, 2D and 3D fractal dimension	Differential diagnosis (malignant vs benign) Treatment response
Statistical features	*1st order - histogram features*: They derive from image histograms (i.e., graphical representation of the intensity distribution of an image) and they describe the distribution of the intensity within the segmentation.	Intensity: minimum, maximum, median, mean, percentiles Kurtosis: magnitude of pixel distribution. It provides a measure of the weight of the histogram tails with respect to a normal distribution. Skewness: asymmetry of the histogram around its mean Entropy: irregularity of the structure. High values correlate with high heterogeneity	Prognosis prediction: locoregional control Survival correlation: overall survival
	2nd order – texture features: They describe the statistical relationship between pixels or voxel to characterize the heterogeneity of the lesion from the segmentation performed for the volume extractions. • Gray-level co-occurrence matrix (GLCM) is the most frequently used. It defines the distribution of concurrent or repeated pixels in the image. • The neighbouring grey-tone difference matrix (NGTDM) uses the intensity values of a neighbourhood instead of one pixel to represent how similar or dissimilar pixel intensities are within a neighbourhood. Other: neighbouring grey level dependence matrix	GLCM: correlation, cluster, contrast, energy NGTDM: complexity, texture strength	
Model-based features	Higher order features are usually based on matrices that consider relationships between three or more pixels or voxels.	Nosologic maps: the spatial registration of the image biomarkers obtained voxel by voxel conforms parametric maps to obtain nosological images that represent different biological behaviours	Prognosis prediction Response prediction

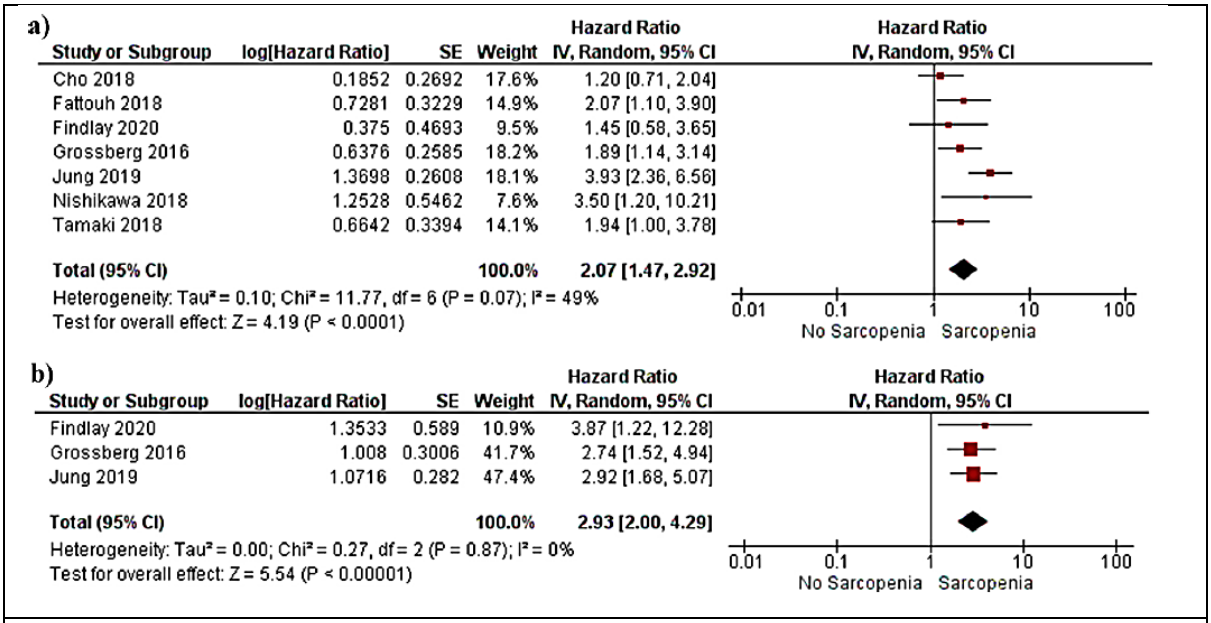
In head and neck cancer the field of radiomics is constantly developing, targeting personalized treatment. Using PET/PET CT biomarkers for patient treatment individualization and response prediction seems promising and literature shows that macroscopic changes in medical images (whether structural or functional) are correlated with biologic and biochemical changes within a tumour [43]. However, there has been

spotted lack of stability and generalization in ongoing research, with the specific study conditions and the authors' choices still influencing considerably the results and although PET radiomics is a promising field, the number of patients in most publications remains inadequate, with very few papers perform in-depth validations [41]. Future research should be directed at overcoming the limitations outlined, mainly regarding sample size, uniformization and standardization of radiomics workflow and subsequent generalization of results, along with optimization of technical issues (e.g., dental artifacts) [40], [41].

1.B.3 The role of sarcopenia, myosteatosis, frailty and nutritional status in patients with head and neck cancer

1.B.3.1 Sarcopenia

The prognostic role of sarcopenia in patients with head and neck cancer is well studied in recent years, and meta-analyses have already been conducted. A meta-analysis of seven studies and 1059 patients where skeletal muscle cross sectional area was evaluated at gold-standard anatomical level of L3 [44] concluded that CT-defined sarcopenia is independently associated with reduced overall survival in patients with HNC and holds a clinically meaningful prognostic value. Forest plots of hazard ratio in subgroup analyses for patients with versus patients without sarcopenia, at pre-treatment and post-treatment time points, for reduced overall survival are shown in Figure 6 [44]. However, given the studies' variation in skeletal muscle index (SMI) threshold values applied and ethnicity, the meta-analysis' authors highlighted the need for consensus regarding sarcopenia assessment and definitions in order to support body composition assessment as a clinically meaningful prognostic tool into practice.



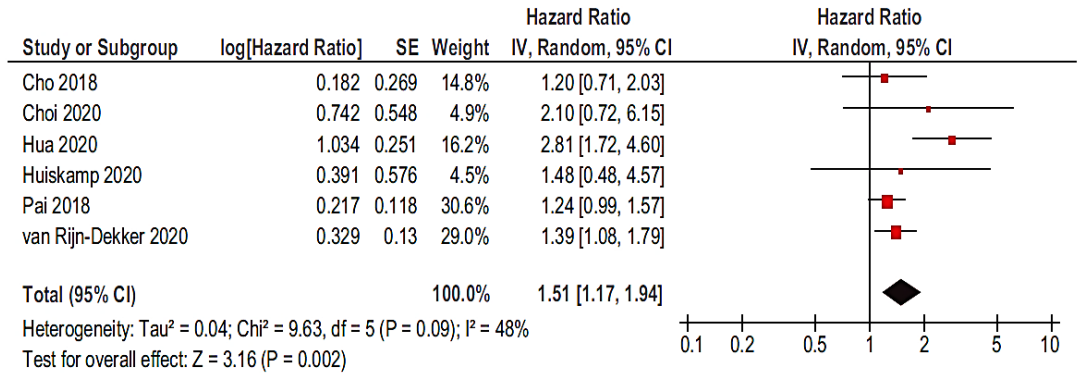
(a) pre-treatment and (b) post-treatment time-points
CI, confidence interval; HR, hazard ratio; SE, standard error

Figure 6: Forest plots of hazard ratio in subgroup analyses for patients with versus patients without sarcopenia for reduced overall survival

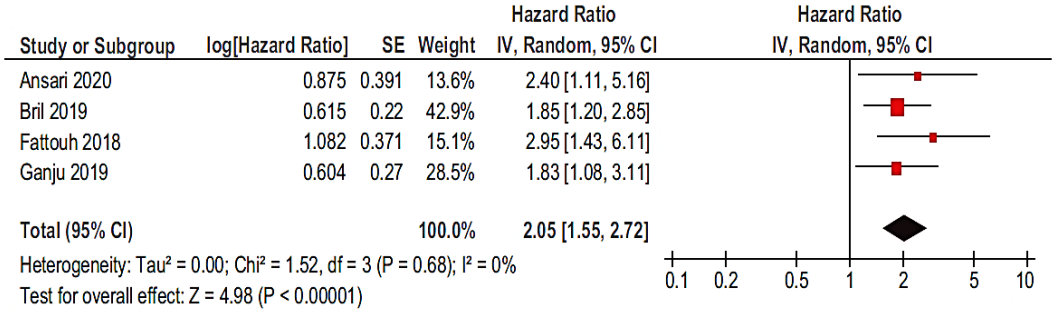
Another meta-analysis included 27 studies with a total of 7704 patients with different HNSCCs [45]. This meta-analysis included both patients treated with definitive chemotherapy and/or radiation, and patients surgically treated with or without adjuvant

chemoradiotherapy. Sarcopenia was associated with lower overall survival (OS) and with occurrence of severe postoperative complications and predicted disease-free survival (DFS) as well. Forest plots of reported hazard ratios of sarcopenia for different treatment modalities used and for different endpoints are shown in Figure 7 [45].

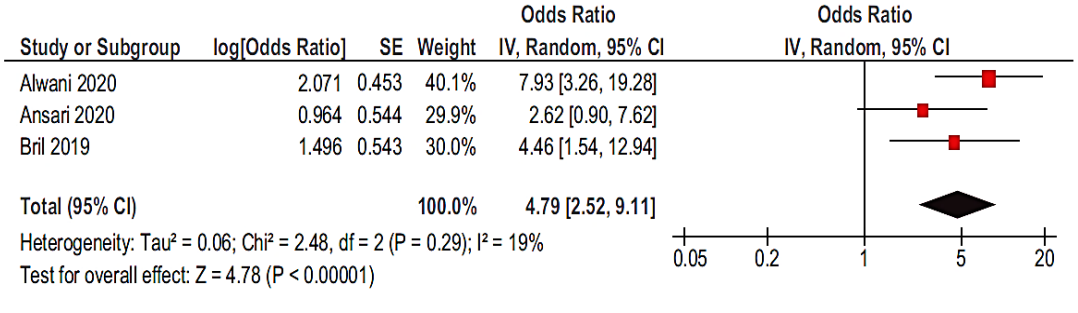
a) Adjusted HRs of sarcopenia regarding OS in patients treated by curative radio-chemotherapy.



b) Adjusted HRs of sarcopenia regarding OS in patients treated by surgery +/- adjuvant radio-chemotherapy.



c) Reported HRs of sarcopenia regarding to occurrence of severe postoperative complications (three or more points according to the Clavien–Dindo classification).



d) Adjusted HRs of sarcopenia relating to DFS

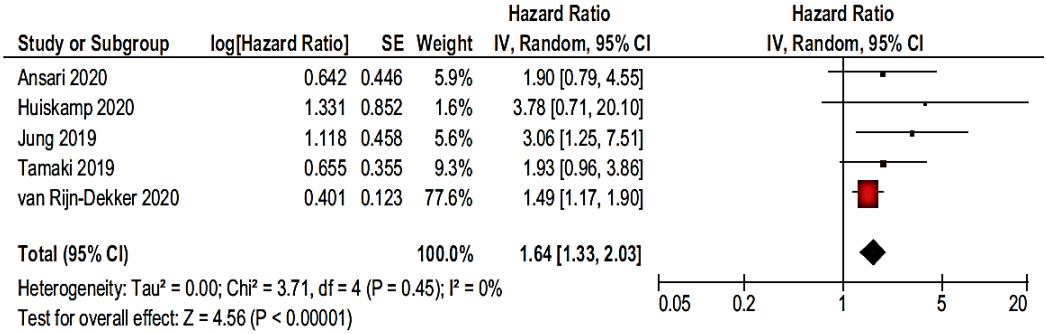


Figure 7: Forest plots of reported HRs of sarcopenia for different treatment modalities used and for different endpoints

Another meta-analysis of 3,461 patients [46] included 11 studies with measures of body composition not limited at L3 (3 studies derived L3 values from equations using measures taken at C3, and 1 study measured at the second thoracic vertebra (T2)). Pre-treatment sarcopenia was independently associated with reduced: overall survival OS (Figure 8 [46]), 3-year OS, disease-free survival, prolonged radiotherapy breaks, and chemotherapy-related toxicities. However, the studies' heterogeneity in HNC diagnosis, ethnicity, definition of sarcopenia, CT level of evaluation, and skeletal muscle index threshold value, led to very low certainty of evidence.

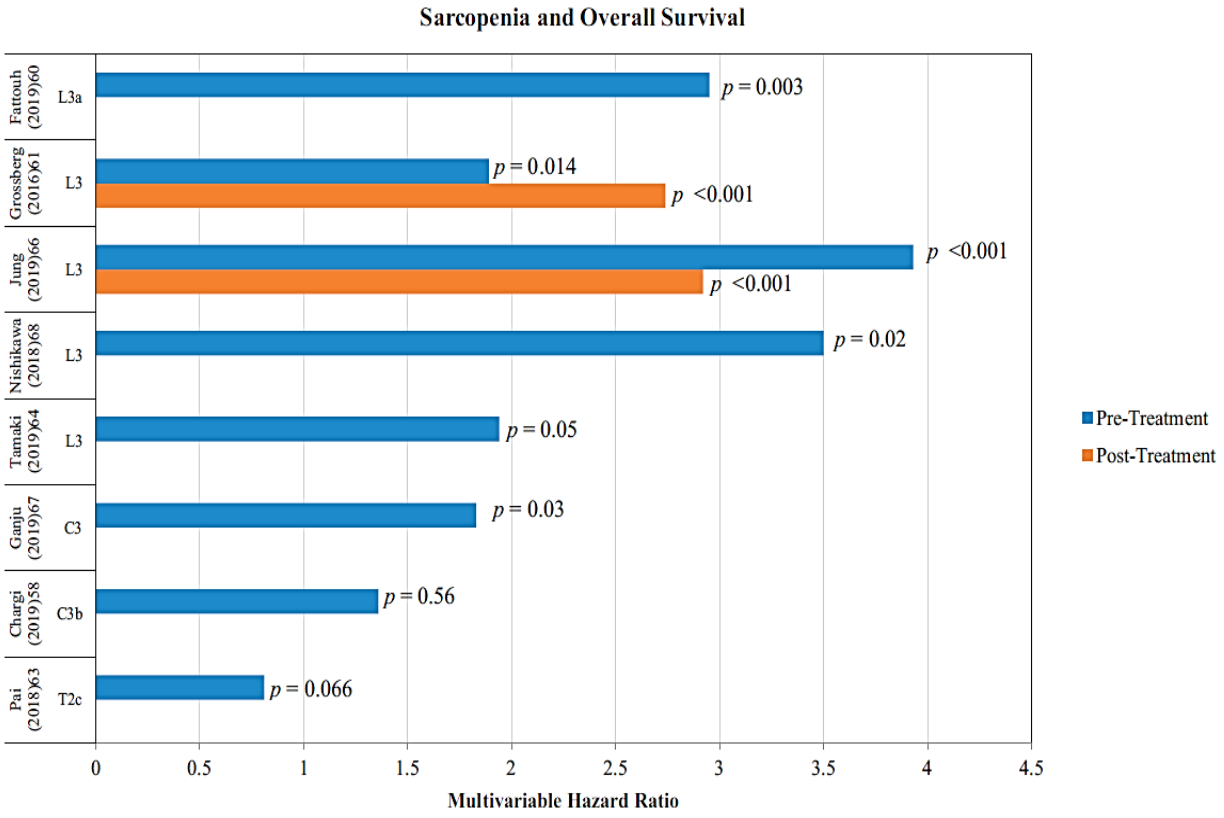
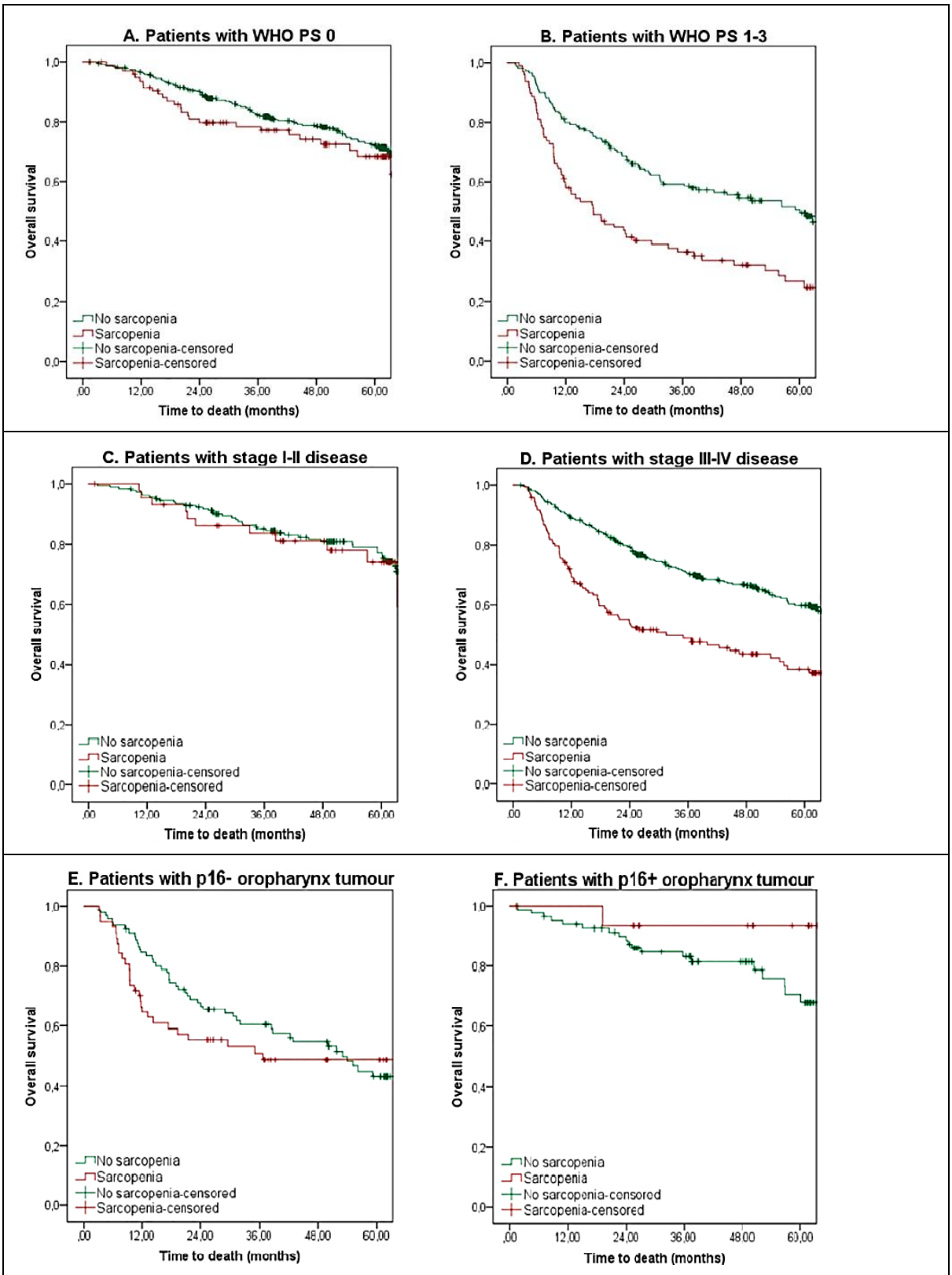


Figure 8: Multivariable hazard ratio for predictive value of sarcopenia on overall survival at pre- and post- treatment time points (muscle status evaluation was undertaken at L3, C3, or T2)

Furthermore, a large cohort study of 750 head and neck cancer patients, treated with definitive (chemo)radiotherapy, that used skeletal muscles at level C3 in order to assess sarcopenia, confirmed that sarcopenia is an independent adverse prognostic factor for OS and DFS, especially in patients with worse World Health Organization Performance Status (WHO PS 1-3), or locally advanced disease, (stage III–IV) [47]. Apart from worse survival outcomes the authors found in multivariable association models, that sarcopenia is associated with physician-rated xerostomia six months after treatment (OR 1.65, p = 0.027) and physician-rated dysphagia six and twelve months after treatment (OR 2.02, p = 0.012 and 2.51, p = 0.003, respectively). Interestingly, the study also showed that in oropharyngeal cancer patients, survival was more determined by p16 status than by sarcopenia. Figure 9 [47] shows Kaplan-Meier curves of OS of the aforementioned study for different patient subgroups. In another study that investigated the prevalence and impact of sarcopenia on DFS and OS in advanced oropharyngeal cancer, see Table 4 [48], sarcopenia was associated with increased mortality and recurrence but was not statistically significant in survival models.



Patients grouped by WHO PS (WHO PS 0, p = 0.154 (A) and WHO PS 1–3, p < 0.001 (B)), stage of disease (stage I–II, p = 0.532 (C) and stage III–IV, p < 0.001 (D)), and p16 status in patients with oropharyngeal cancer (p16 negative, p = 0.541 (E) and p16 positive, p = 0.150 (F)).

Figure 9: Kaplan-Meier curves of overall survival in head and neck cancer patients treated with radiotherapy

Table 4: Multivariable Cox proportional hazards model of OS and DFS OS in advanced stage oropharyngeal cancer

		Hazard Ratio	95% CI	p-value
OS	HPV 16	0.463	(0.235-0.909)	0.25
	Sarcopenia	1.943	(0.999-3.779)	0.50
DFS	HPV 16	0.403	(0.201-0.810)	0.11
	Sarcopenia	1.926	(0.961-3.862)	0.65

Finally, one of the few studies that have investigated the impact on cost, while focusing on CT-defined sarcopenia, suggests that, compared to patients who were never sarcopenic, the mean cost of unplanned admissions was higher for patients who were sarcopenic either pre-treatment or post-treatment, as well as for those who became sarcopenic during care [49]. Unplanned admissions usually occur due to increased susceptibility to treatment toxicities, malnutrition, dehydration, and psychosocial impact. Therefore, understanding the impact CT-defined sarcopenia has on outcomes for these patients, holds possible important implications regarding nutrition interventions and individualized care.

1.B.3.2 Myosteatosis

There are very few studies assessing myosteatosis on survival outcomes for patients with head and neck cancer. Myosteatosis is usually assessed through calculation of mean muscle attenuation on CT scan (MACT) for the entire L3 muscle area [50]. MACT threshold values were defined for both sexes according to body mass index (BMI) using optimal stratification based on log-rank statistics to best separate patients with respect to time to death (Table 5 [50]).

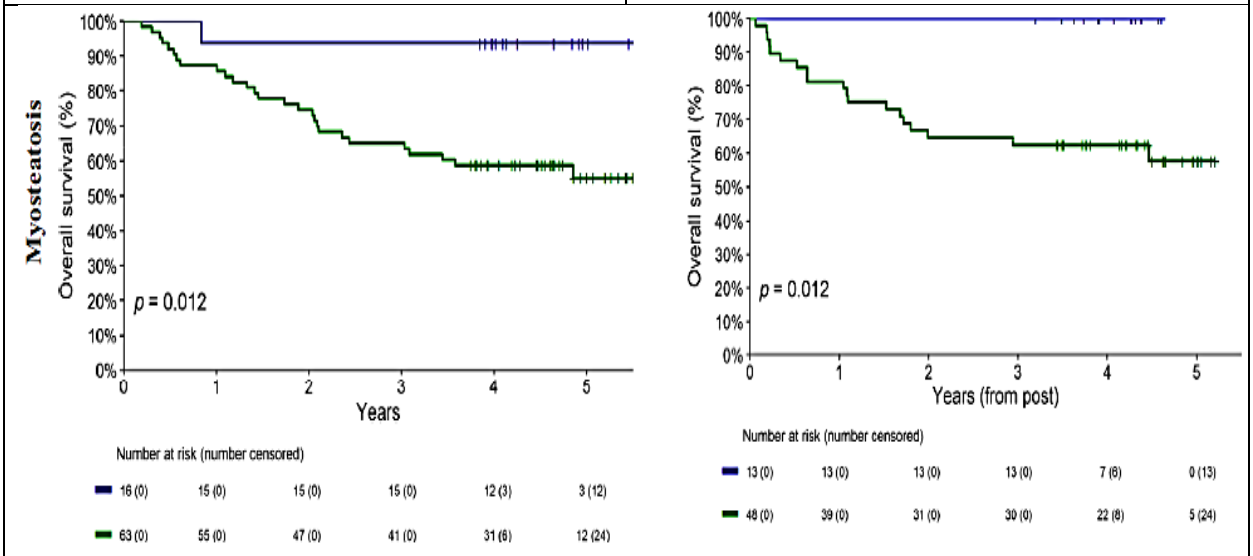
Table 5: MACT Threshold values significantly associated with low survival

BMI Category (kg/m2)	MACT (HU)	
	Men	Women
Underweight (< 20.0)	<41	<41
Normal weight (20.0 to 24.9)	<41	<41
Overweight (25.0 to 29.9)	<33	<33
Obese (≥ 30.0)	<33	<33

In one of those few studies [51] the very high prevalence of pre-existing myosteatosis (in over 90% of participants) prevented any meaningful statistical comparison with the very small non-myosteatotic group, and therefore no significant association with outcomes was observed. However, another retrospective observational study [49], found that CT-defined myosteatosis holds clinically meaningful prognostic value and recommended muscle status evaluation in routine clinical practice, when treating patients with head and neck cancer. In the aforementioned study, pre-treatment myosteatosis was significantly associated with overall survival both in univariate and multivariate analysis (adjusted for possible confounders including gender, age, TNM stage, treatment modality, body mass

index category, tobacco use, alcohol use and human papilloma virus status). The corresponding survival curves are shown in Figure 10 [49] .

Pre-treatment	Post-treatment



green curve: patients with myosteatosis, blue curve: patients without myosteatosis

Figure 10: Kaplan–Meier survival estimates of overall survival with log-rank comparisons of patients with and without myosteatosis at pre-treatment and post-treatment

1.B.3.3 Frailty

Frailty has been found as an important determinant of many health outcomes across various surgical specialties and is an emerging predictor of outcome in elderly HNC patients. Although functional and cognitive impairment, depressive symptoms and social isolation had been associated with high risk of worse prognosis in older patients with head and neck cancer, since 2017 no studies reported association between frailty and adverse health outcomes [52]. However, the current literature demonstrates the utility of frailty as a predictor of perioperative mortality and morbidity, with recent studies supporting a significant association between frailty and perioperative outcomes, length of hospital stay, readmission rate, and likelihood of discharge to short-term or skilled nursing facilities [28]. In a prospective study with 274 patients recruited [53] , frailty was a predictor of both type and severity of complications and an independent predictor of length of hospital stay. Frailty and functional assessment can help surgeons identify patients at risk of adverse postoperative outcomes, still further research is needed to develop frailty screening measures in order to risk-stratify patients and optimize modifiable factors preoperatively.

1.B.3.4 Nutritional status

The Nutritional Risk Screening-2002 (NRS-2002) and Patient-Generated Subjective Global Assessment (PG-SGA) are the most common tools used for nutritional assessment [54] and high nutritional risk according to the NRS-2202 and worse nutritional status according to the PG-SGA are positively associated with a longer hospital stay and mortality. The PG-SGA is a subjective nutritional assessment tool used in oncology and other chronic catabolic conditions, including questions about symptoms of nutritional

impact and recent weight loss. The PG-SGA allows to classify patients as well-nourished (A) or either moderately (B) or severely (C) malnourished.

A retrospective, observational study of 277 patients who had completed radiotherapy (RT) or chemoradiotherapy (CRT) of curative intent for HNC aimed to describe body composition profile and examine the impact of nutritional status as well as independently and concurrently occurring body composition features on overall survival, treatment completion, unplanned admissions, and length of stay [55]. PG-SGA was used to determine nutritional status, tissue-density data were derived at the third lumbar vertebra (L3) with sarcopenia and myosteatosis defined by published, sex-specific threshold values stratified by body mass index for skeletal muscle index (cm2/m2) and skeletal muscle radiodensity (SMR, Hounsfield Unit). The prevalence of malnutrition was 24.9% of sarcopenia 52.3%, of myosteatosis 82.3%, and of concurrently occurring sarcopenia and myosteatosis 39.7% (Figure 11[55]).

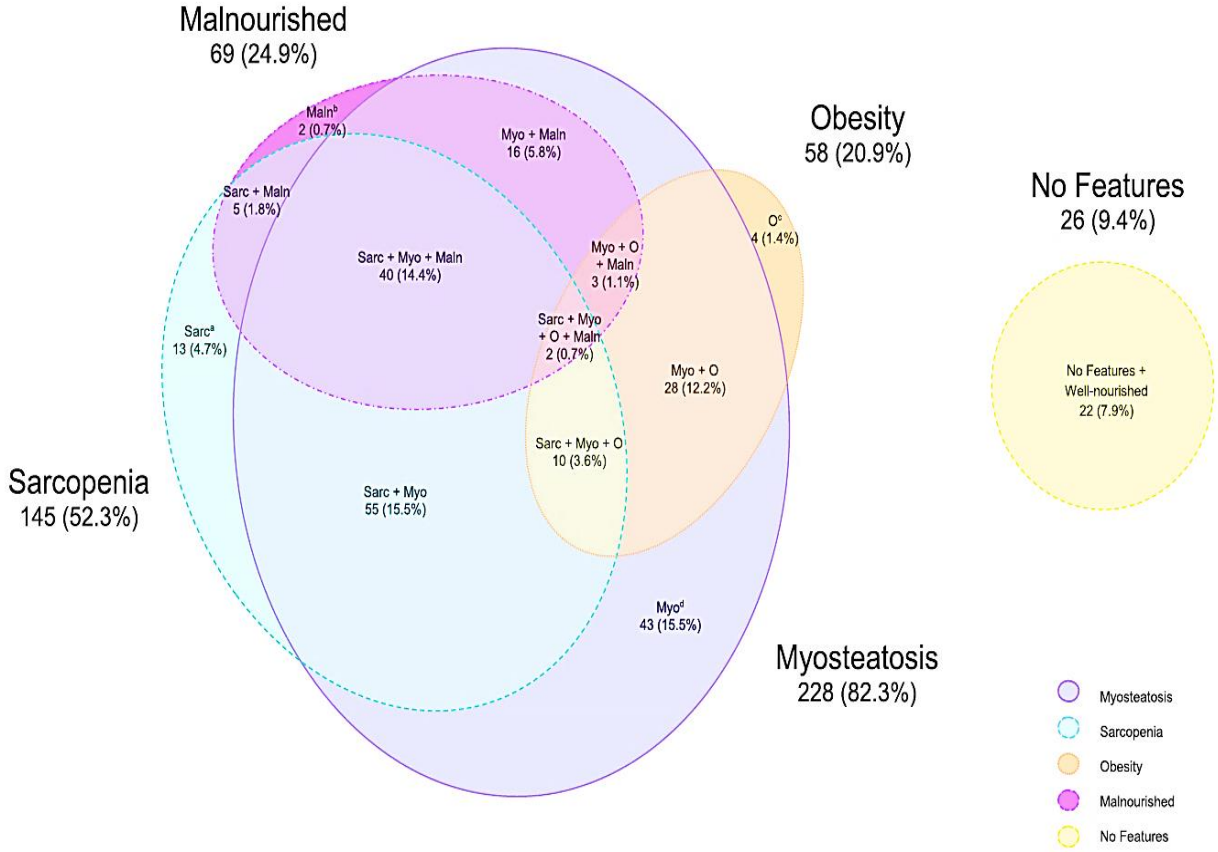


Figure 11: Euler diagram denoting pre-treatment combination of computed tomography-defined body composition features and baseline nutritional status

Malnutrition was found to be a more powerful prognostic indicator than CT-defined skeletal muscle depletion, independently associated with reduced OS in patients undergoing radiotherapy or chemoradiotherapy of curative intent for HNC. Figure 12 [55] shows Kaplan–Meier survival estimates of overall survival with log-rank comparisons for combination of skeletal muscle status features and for nutritional status. Moreover, malnourished patients were more likely to require unplanned hospital admission with 58% of severely malnourished patients vs. 34% of well-nourished patients admitted ($p = 0.021$), Therefore, the authors suggested that CT-defined skeletal muscle depletion studies should also measure nutritional status using validated methods in order to develop more accurate high risk stratification criteria for the complex group of patients with head and neck cancer.

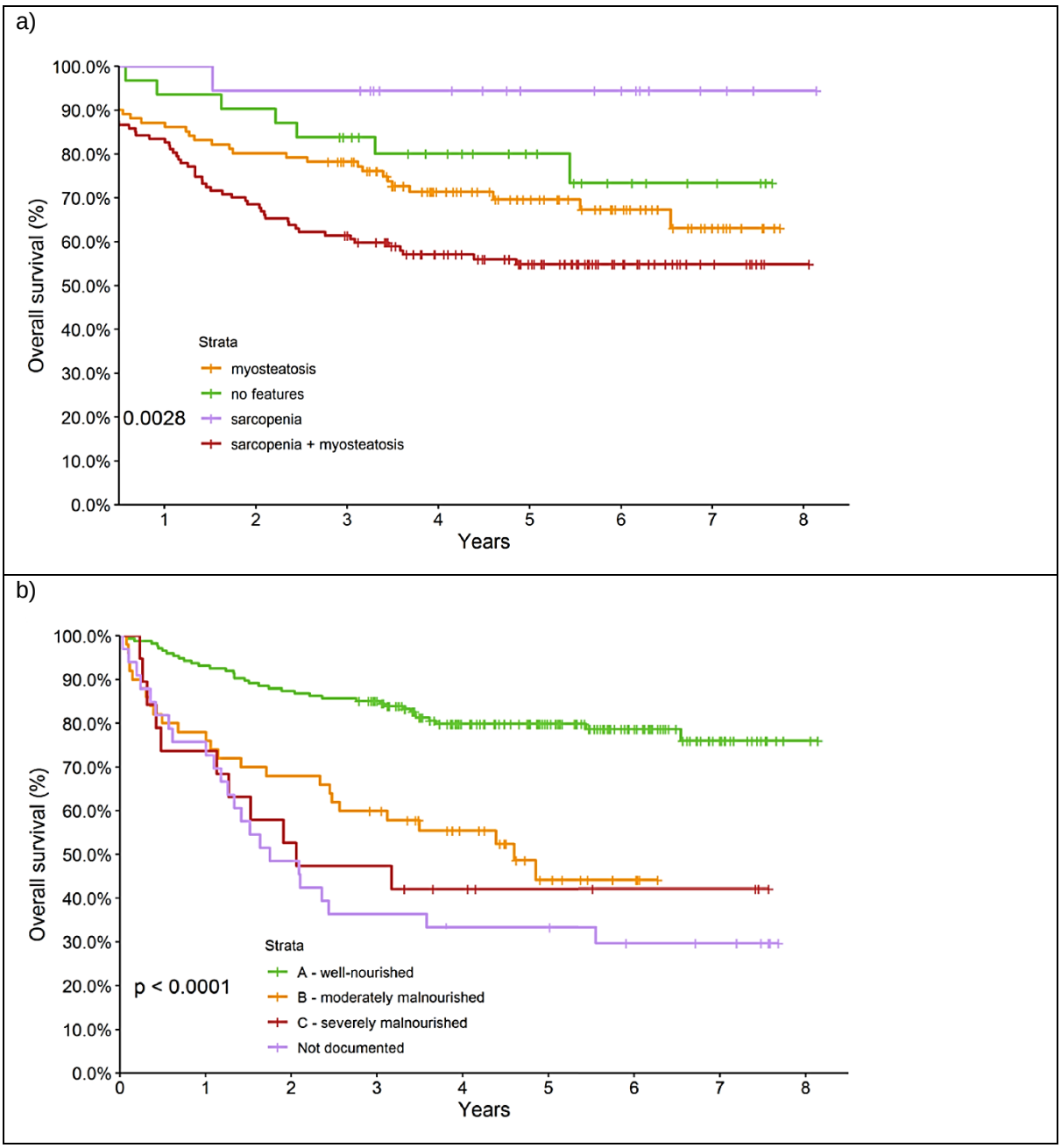


Figure 12: Kaplan–Meier survival estimates of overall survival with log-rank comparisons for combination of skeletal muscle status features (a) and for nutritional status (b)

1.B.4 Reviewing the literature on how sarcopenia is assessed from computed tomography in patients with head and neck cancer and how widespread is each method when addressing sarcopenia's prognostic role

CT allows the evaluation of muscle quality and fatty infiltration [56]. Abdominal CT-imaging at the level of the third lumbar spine vertebra (L3) has been broadly used to assess sarcopenia, as the cross-sectional area (CSA) of the skeletal muscles measured at the level of L3, correlates well with the total-body skeletal muscle mass [57]. In HNC patients though, such scans are rarely available. Before 2016 there was hardly any published literature regarding the effect of sarcopenia in HNC patients, probably because of the absence of a widely available diagnostic tool to assess sarcopenia in those patients. In 2016, a study [58] investigated the feasibility of using head and neck CT imaging in order to assess skeletal muscle mass in HNC patients. The authors compared

muscle CSA at the level of C3 to L3 and correlated skeletal muscle mass assessed on head and neck CT-scans with abdominal CT imaging, concluding that assessment of skeletal muscle mass on head and neck CT-scans is feasible and may be an alternative to abdominal CT-imaging. Therefore, C3-level CT-scans, which are routinely performed in HNC patients, offer a cost-effective and widely available tool to determine sarcopenia, allowing for assessment of sarcopenia in HNC patients without additional imaging.

The literature review was performed by following the PRISMA 2020 (preferred reporting items for systematic reviews and meta-analyses) statement [59] (Figure 6). The bibliographic databases PubMed/MEDLINE [60] and Scopus were searched manually for relevant published studies reporting how sarcopenia is assessed from computed tomography in HNSCC patients, using the keywords: ((computed tomography) AND sarcopenia) AND (head and neck). The eligibility criteria for including studies in the present review were the following: (i) studies reporting the effectiveness of head and neck or thoracic CT images to assess SMM in patients with HNC and/or (ii) studies addressing the prognostic role of sarcopenia in HNSCC patients. Studies were excluded from this review based on the following exclusion criteria: (i) not directly assessing sarcopenia, (ii) cut-offs were determined by optimal stratification of cohort's data according to the outcome of interest (usage of outcome-oriented optimal cut-off methods) (iii) reviews, editorials, commentaries. Prognostic studies where data-oriented stratification methods were used, such as using median or quartiles for cut-off values, were not excluded. Moreover, studies using optimal cut-off values for sarcopenia prediction, were also included. Notably studies using optimal cut-offs previously proposed from different cohorts – studies were included, for evaluation purposes. Collectively, 76 relevant records were retrieved from PubMed (up to 06 April 2023) (Figure 13) and 85 records from Scopus. 85 records were removed before screening 72 duplicates and 13 reviews and letters. After initial screening, 26 titles and abstracts were excluded because they were irrelevant to our study. A total of 50 full-text articles were assessed for eligibility. By applying strict inclusion and exclusion criteria, 31 studies were included in this review (Fig. 14). The basic characteristics of the included studies are summarized in Table 6.

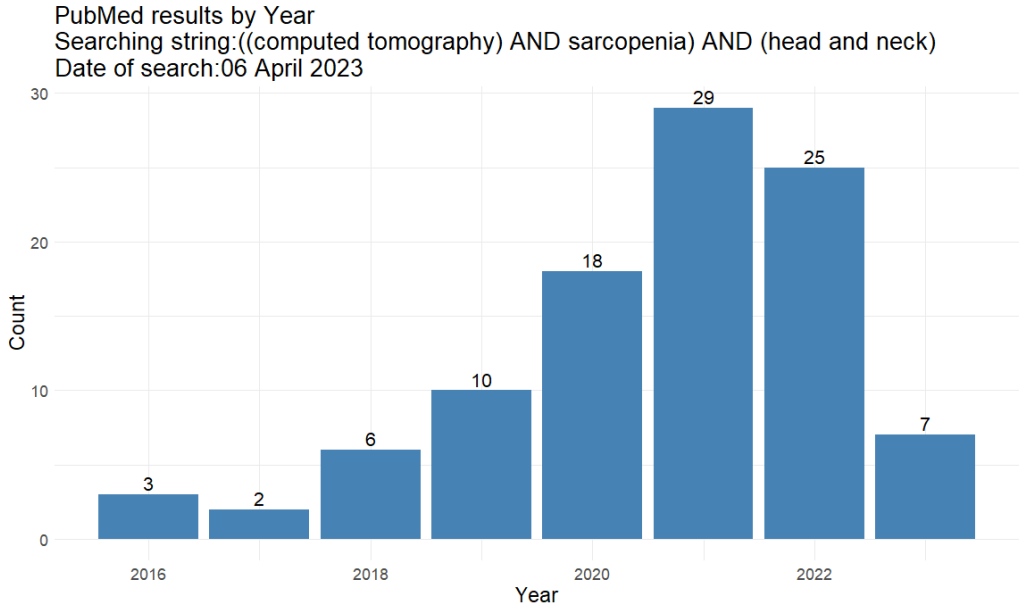


Figure 13: Relevant records retrieved from PubMed by year (date of search: 06 April 2023)

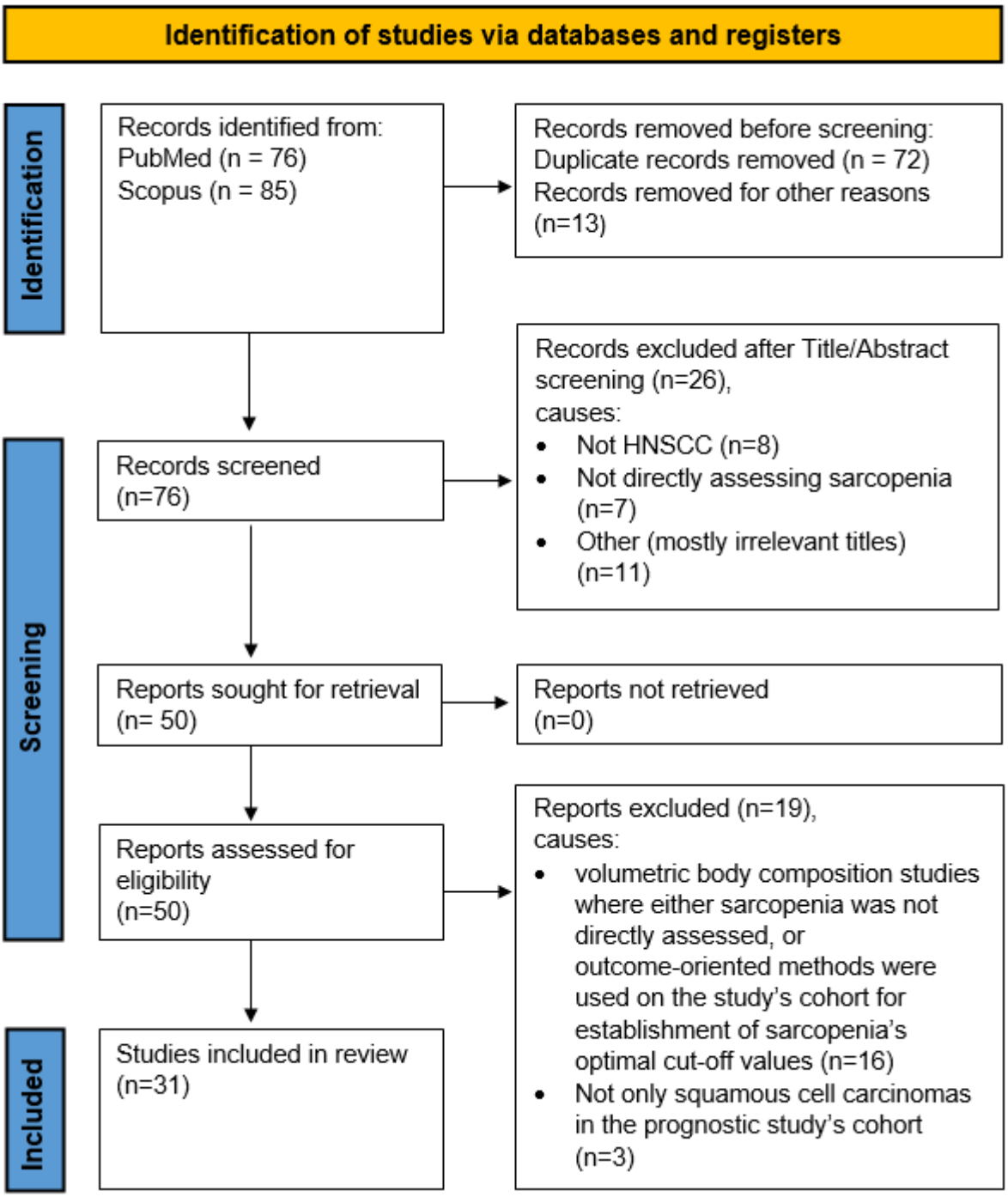


Figure 14: PRISMA 2020 flow diagram

Table 6: Main characteristics of the eligible studies included in the review, that have investigated the feasibility to use alternate to L3 skeletal muscle areas in order to assess sarcopenia in patients with HNSCC

Skeletal muscle area		Studies investigating feasibility	Comparison with gold standard (L3); variable tested for correlation (r, p value)	Supports usage for assessing L3?	Sarcopenia cut-off values and/or formula for L3 prediction
Cervical vertebral level	C2	[61]	SMI (r=0.810, p<0.001)	Yes	Men: 9.3cm^2/m^2, Women: 8cm^2/m^2
			Note: cut-off values and prediction rule were obtained by multivariable analysis and by evaluating the diagnostic performance of these values in the diagnosis of sarcopenia via ROC curve		
	C3	[62]	Predicted L3-SMI (r=0.883, p<0.001)	Yes	L3-CSA = 124.838 + [1.881*C3-CSA] – [24.687*sex] – age + [0.472*Weight] (male:1, female:2)
			Note: The prediction model for estimating L3-CSA in this study's predominantly overweight cohort was found to have better agreement, and specificity than that of [63] suggesting probable better effectiveness in recognizing sarcopenic obesity.		
		[64]	C3-CSA (r=0.810, p<0.001), predicted L3-CSA (r=0.875, p<0.0001)	Yes	L3SMI cut-off for men:<55.0cm^2/m^2, for women:<36.6cm^2/m^2 L3-CSA= −6.310 + 1.845*C3-CSA + 1.101*Weight + 4.923*Sex (female = −1, male = 1)
			Note: X-tile was used for cut-offs establishment, which applies an outcome orientated optimal cut-off method.		
		[65]	SMA (men (r=0.77, p<0.001), women (r=0.80, p<0.001))	Yes	Men: 14cm^2/m^2, Women: 11.1cm^2/m^2
			Note: ROC curves were generated to show the general predictive ability of C3 to predict L3-defined sarcopenia and Youden's Index was used to determine the optimal C3 cut-off value for predicting sarcopenia		
		[66]	Predicted L3-CSA (r=0.86, p<0.001)	Yes, but	The [63] formula for L3 prediction was used. Each patient was classified as sarcopenic or not by applying the sex and BMI-specific threshold values at L3 [50]
			Note: Sarcopenia was diagnosed in 26%-(L3), 45%-(C3), with weak agreement (sensitivity 79.2%, specificity 66.7%). The study highlighted the limitation of applying predefined prediction formulas on different populations.		

		[67]	C3-CSA (r=0.75, p<0.01), predicted L3-CSA (r=0.82, p<0.01)	Yes, but	The [63] formula for L3 prediction was used
			Note: There is moderate agreement in the identification of patients with low SMM based on the estimated lumbar SMI (based on measurement at C3) and actual LSMI		
		[68]	C3-SMM in non-sarcopenic patients (r = 0.876, p<0.001), while in sarcopenic patients (r = 0.381, p=0.003). Predicted L3-SMM in non-sarcopenic patients (r>0.9, p<0.001), whereas in sarcopenic patients (r = 0.7633, p<0.0001).	No	Predicted L3 = 45.9183 + 0.9736*C3-PVM + 1.2863*Weight – 0.4414*Age – 18.2159*Sex (male:0, female:1)
			Note: correlation between L3 and C3 SMMs was weak in sarcopenic patients and the prediction model showed poor diagnostic accuracy. Therefore, C3 SMM may not be a strong predictor for L3 SMM in sarcopenic HNC patients.		
		[69]	C3-SMM (r=0.421, p<0.001), Predicted L3-SMM(r=0.721, p<0.001)	Yes	56.3cm^2 L3-SMM= 81.059 + 0.874*C3-SMM + 0.956*Weight − 28.127* Sex − 0.257*Age
			Note: supports usage of prediction model including the strongest predictive factors (sex, age, weight, and C3-CSA) as it significantly increased the L3-CSA correlation power. Median C3- SMM value was used as cut-off.		
		[61]	SMI (r=0.877, p<0.001)	Yes	Men: 9.3cm^2/m^2, Women: 6.3cm^2/m^2
			Note: cut-off values and prediction rule were obtained by multivariable analysis and by evaluating the diagnostic performance of these values in the diagnosis of sarcopenia via ROC curve		
		[63]	C3-CSA (r=0.785, p<0.001), Predicted L3-CSA (r=0.891, p<0.001)	Yes	L3-CSA= 27.304 + 1.363*C3-CSA -0.671*Age + 0.640*Weight + 26.442*Sex (female:1, male:2)
	C4	[61]	SMI (r=0.827, p<0.001)	Yes	Men: 10.8cm^2/m^2 Women: 9.5cm^2/m^2
			Note: cut-off values and prediction rule were obtained by multivariable analysis and by evaluating the diagnostic performance of these values in the diagnosis of sarcopenia via ROC curve		

Thoracic vertebral level	T2	[70]	Predicted L3-CSA (r=0.796, p<0.001)	Yes	L3-CSA = 174.15+[0.212*T2-CSA] - [40.032*Sex] - [0.928*Age] + [0.286*Weight] (male:1, female:2)
	T4	[71]	Muscle CSA (r=0.791, p<0.05).	Yes	L3-CSA = 34.48 + 0.78 * T4- CSA
			Note: measurements at the level of T4 can be an alternative in patients with extensive localized disease or post-treatment necks		
	T12	[72]	Muscle CSA (r= 0.915 95%CI [0.886–0.937], p<0.05)	Yes	L3-CSA = 14.143 + 0.779*T12-CSA - 0.212*Age + 0.502*Weight + 13.763*Sex (female:1, male:2)
Masticatory (Masseter, pterygoid, temporalis)		[73]	Masseter muscle volume corelation with L3-CSA (r=0.531, p<0.001)	No	patients present in the lowest quartile of MCSA for their specific gender as "low MSMI"
		[74]	Masticatory SMI (r=0.901, p<0.001)	Yes	MSMI of <5.5cm^2/m^2 was an independent predictor of sarcopenia (hazard ratio = 5.37, p < 0.001) L3SMI= 7.21*MSMI + 7.56
			Note: ROC curve analysis was used to assess the ability of the Masticatory SMI to identify sarcopenia, and Cox logistic regression was used to identify predictors of sarcopenia		
Infrahyoid		[75]	SMI (r=0.434, p<0.001)	No	16.88 cm^2/m^2
			Note: cut- off value according to ROC curve analysis using Youden's index by referencing the overall survival (OS). L3SMI and IHSMI were moderately correlated. However, IHSMI might be a good predictor for OS.		
Abbreviations: SMI, skeletal mass index; SMA, skeletal muscle area; SMM, skeletal muscle mass; r, Pearson correlation coefficient; CSA, cross-sectional area; CI, confidence interval; IHSMI, infrahyoid skeletal muscle index; L3SMI, L3 skeletal muscle index; OS, overall survival; ROC, receiver operating characteristic; PVM, paravertebral muscle; MSMI, masticatory skeletal muscle index;					
Overall note: Several studies have suggested formulas and cut-off values related to sarcopenia. However, the results varied considerably, possibly due to different races, regions, age groups and disease conditions including stage (all stages vs locally advanced carcinomas), primary site and virus relation (e.g., HPV (+) oropharyngeal carcinomas).					

Table 7: Main characteristics of the eligible studies included in the review that have addressed the prognostic role of sarcopenia in patients with HNSCC without using outcome-oriented methods for sarcopenia cut-off values establishment

Skeletal muscle area	Studies in total per area	Studies on the prognostic role of sarcopenia	Number of patients (N); Primary; Stage; Outcome	Sarcopenia assessment and patient stratification	Survival statistic, p value
.L3	9	[51]	N=101; All sites (mostly p16 positive OPC); All stages; OS	Sarcopenia was defined according to [50] : for females: SMI<41cm^2/m^2 and for males: SMI<43cm^2/m^2 for BMI ≤ 24.9kg/m^2 or SMI<53 m^2/m^2 for BMI ≥ 25 kg/m^2. Both baseline and post-treatment sarcopenia were assessed.	5-year OS favoured those without post-treatment sarcopenia (HR=0.37, 95%CI [0.16-0.88], p=0.06). No significant differences found in OS regarding the presence of baseline sarcopenia.
		[76]	N=216; All sites; Stages II-IV; OS, DFS	Cut-off values for men: SMI<43.3cm^2/m^2 and for women: SMI<33.09cm^2/m^2 (lowest gender specific quartile values of our population) according to [47]	3-year OS was 75% versus 82% (p=0.1) and 3-year DFS was 70% versus 85% (p=0.00015) for sarcopenic and non-sarcopenic patients, respectively. Pre-treatment sarcopenia was an independent negative prognostic factor for DFS (HR=2.174, p=0.0001).
		[77]	N=190 (patients aged ≥65 years who underwent curative surgery); All sites; All stages; OS, DFS, CSS, LRFS	Sarcopenia was defined as SMI<52.4cm^2/m^2 for men and SMI<38.5cm^2/m^2 for women based on [78]	Patients with sarcopenia before treatment had about a 4.5-fold increased risk of overall recurrence or death. 5-year OS rates of patients without and with pre-treatment sarcopenia were 79.7% and 20.4% respectively (p<0.001). 5-year DFS rates of patients without and with pre-treatment sarcopenia were 82.2% and 26.0%, respectively (p<0.001). Sarcopenia was also the significant factor of cause-specific death (HR=5.33, 95%CI [3.05–9.31], p< 0.001) and local control (HR=5.89, 95%CI [2.94–11.79], p<0.001). In multivariate analyses, sarcopenia remained strongly associated with OS and DFS (p<0.001).
		[79]	N=113; All sites; All stages; OS	Sarcopenia was defined based on sex specific cut-off values established by [78]	Sarcopenic patients had poorer OS compared to non-sarcopenic (Log-rank p=0.004). When stratified by BMI group, OS in sarcopenic patients remained significantly poorer, regardless of BMI group prior to treatment.

		[80]	N=258; All sites; Stages III/IV; OS, DFS	Sarcopenia was defined based on sex specific cut-off values established by [78]	Sarcopenia was significantly associated with DFS and OS (all p<0.05). In multivariable analysis both pre-treatment and post-treatment sarcopenia remained independent variables predictive of DFS (pre-treatment sarcopenia: HR=3.06, 95%CI [1.25-7.54], p=0.015; post-treatment sarcopenia HR=3.34, 95%CI [1.70-6.55], p<0.001) and OS (pre-treatment sarcopenia: HR=3.93, 95%CI [2.36-6.56], p<0.001; post-treatment sarcopenia HR=2.92, 95%CI [1.68-5.07], p<0.001).
		[48]	N=113; Oropharynx, Stages II-IVC; OS, DFS	Sarcopenia was defined using SMI thresholds proposed by [50].	Log-rank tests of differences in survival distributions did not reveal differences across DFS (p=0.065) but did demonstrate a statistically significant difference with OS (p=0.049). However, sarcopenia was not a statistically significant predictor of OS (HR=1.925, 95%CI [0.993-3.735], p=0.053) or DFS (HR=1.901, 95%CI [0.950-3.802], p=0.069) on univariable analysis.
		[81]	N=221; All sites; Stages III/IV; OS, PFS	Sarcopenia was defined as SMI<49cm^2/m^2 for men and SMI<31cm^2/m^2 for women based on previous studies of the same ethnicity (Korean). [82], [83]	Sarcopenic patients showed poorer OS than non-sarcopenic patients (3-year OS: 62 vs. 76%, p=0.037), but PFS rates were not significantly different between the 2 groups (3-year PFS: 46.6 vs. 55.6%, p=0.187).
		[84]	N=158; Larynx and Oropharynx; All stages; OS, PFS	Sarcopenia was defined based on sex specific cut-off values established by [78]	Sarcopenia was not independently predictive for increased risk for overall death and disease progression
		[85]	N=190 (oropharyngeal=139, non-oropharyngeal=59), All sites; All stages; OS, CSS, LRFS	Sarcopenia was defined based on sex specific cut-off values established by [78]	Pre-treatment sarcopenia was significantly associated with shorter OS (HR=1.92, 95%CI [1.19–3.11], p=0.007) and CSS (HR=1.87, 95%CI [1.03–3.36], p=0.03). No significant difference in LRFS was observed (HR=1.38, 95%CI [0.66–2.89], p=0.34)

			Note: Separate analysis regarding primary site was performed finding that sarcopenia was associated with a decrease in OS (HR=1.89, 95%CI [0.94–4.23], p=0.09) and CSS (HR=2.85, 95%CI [1.20–7.20], p=0.02) in patients with non-oropharyngeal HNSCC, but not in those with oropharyngeal carcinomas		
C3	8	[86]	N=426; OSCC; All stages; OS	Predicted L3 [63], previous cut-off established in the literature [87]	Sarcopenia did not seem to cause a statistically significant reduction in OS in patients with OSCC (HR=0.996, 95%CI [0.732-1.354], p=0.979), however,
			Note: sarcopenic obesity showed a meaningful negative prognostic impact on OS (HR=0.985, 95%CI [0.424-2.286], p=0.972)		
		[65]	N=536; All sites; All stages; OS	Optimal C3 cut-off value for predicting sarcopenia (C3-SMI cut-off for men: 14cm^2/m^2 and for women: 11.1cm^2/m^2)	C3 sarcopenia was independently associated with reduced overall survival in men (HR = 2.63; 95%CI [1.79, 3.85], p<0.0001) but not women (HR = 1.18, 95% CI [0.76, 1.85], p=0.46)
		[88]	N=300, Oropharynx, Supraglottic Larynx, Hypopharynx; stages III/IV; DFS	Predicted L3 [63], previous cut-off established in the literature [89]	As per cut of criteria used nearly 91% of the patient cohort were sarcopenic. Sarcopenic patients had a worse DFS
		[90]	N=164; All sites; All stages; OS, DFS, LRFS	Predicted L3 [63], previous cut-off established in the literature [87]	The sarcopenia group had poorer 3-year OS (73.3% vs. 94.7%, p<0.01). There were no significant differences in 3-year DFS (p=0.084) or 3-year LRFS (p=0.34). In the multivariate analysis, sarcopenia (HR=2.95, 95%CI [1.34–6.49], p<0.01) was significantly associated with poor OS.
		[91]	N=,174 OSCC; Stages III-IVB; OS and DFS	Predicted L3 [63], cut-off values for sarcopenia were set at the lowest tertile for SMI	The 5-year OS rate was 54.0% in the sarcopenic group and 79.0% in the non-sarcopenic group (p=0.001); the corresponding 5-year DFS rates were 48.0% and 78.3%, respectively (p=0.006)
		[47]	N=750; All sites; All Stages; OS and DFS	Predicted L3 [63], cut-off values for sarcopenia were set at SMI according to lowest gender specific quartile	Three-year OS and DFS in sarcopenic patients were 56% and 48% versus 75% and 69% in non-sarcopenic patients, respectively (both p<0.001). When stratified by stage of disease significant difference was found only in advanced stages (stage I–II, p=0.532 and stage III–IV, p<0.001).

		[69]	N=305; All sites; Stages III/IV; OS	Median C3-SMM was used as cut-off	5-year OS rates of low and high C3 SMM were 46.3% and 87.6%, respectively (p<0.001). Multivariate analysis showed that C3 SMM remained independent variable predictive of OS (p<0.001)
		[92]	N=246; All sites; Stages III-IVB; OS and PFS	Predicted L3 [63], cut-off values for sarcopenia according to gender specific SMI thresholds proposed by [50]	While sarcopenic patients had worse survival outcomes overall, this was driven by patients without p16-positive oropharyngeal cancers. In p16-positive oropharynx patients, there was no difference in either OS (p=0.82) or PFS (p=0.38) in sarcopenic compared to non-sarcopenic patients. In all other patients, the difference in OS (p=0.01) and PFS (p=0.02) remained significant, with the estimated OS at 3 years 71.2% and 53.2%, and estimated PFS at 3 years 78.1% and 56.3% in patients without and with sarcopenia, respectively.
Masseter	2	[73]	N=99; All sites; All stages (mostly III/IV); OS	Gender specific low quartile of MCSA	significant difference in OS, p=0.015
		[93]	N=111; All sites; All stages; OS	sarcopenia defined using as cut-off the gender based mean MCSA	significantly associated with worse OS, p=0.038
T2	1	[70]	N=111; All sites; All stages; cancer-specific survival (CSS)	Predicted L3	no significant difference in 5-year CSS, p=0.191
				Low quartile T2-SMI	significantly worse 5-year CSS, p=0.003
Abbreviations: SMI, skeletal mass index; CSS, cancer-specific survival; OS, overall survival; MCSA, masseter cross-sectional area; SMM, skeletal muscle mass; OSCC, oral squamous cell carcinoma; PCF, pharyngocutaneous fistula; TL, total laryngectomy; DFS, disease free survival; PFS, progression free survival; OPC, oropharyngeal carcinoma; LRFS, locoregional recurrence-free survival;					

1.B.5 Application of radiomics and machine learning in head and neck cancer

In general machine learning (ML) could be described as computational algorithms using data to improve performance or make accurate predictions. In the field of head and neck cancer, artificial intelligence (AI)-based prediction models have been created for both oncologic outcomes, treatment toxicity, and pathological findings (Figure 15) [94]. Especially in the field of Radiation Oncology machine learning methods have been applied in auto-segmentation, treatment planning optimization, and prediction of oncological and toxicity outcomes [95].

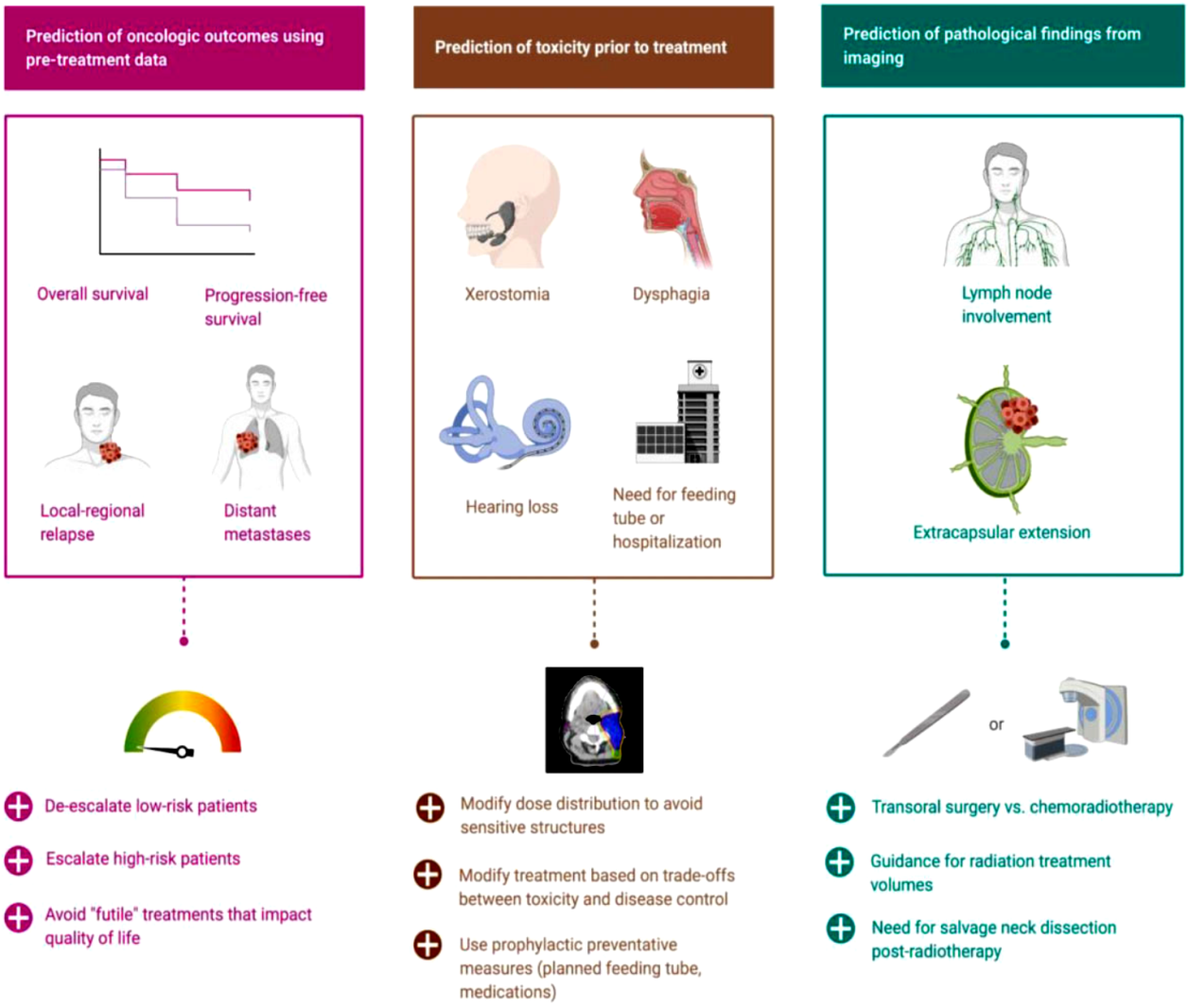


Figure 15: Applications of AI – based prediction models in head and neck oncology

In patients with HNSCCs, treatment choices are made aiming to achieve disease control while maintaining an acceptable treatment toxicity. The ability to accurately predict treatment outcomes through ML models, allows for personalized treatment intensity choices [94]. However, precision medicine using radiomics and artificial intelligence is heavily dependent on the quality, robustness, and generalizability (model's ability to perform well in new unseen data) of generated prediction models [96]. Critical challenges regarding model generalizability are false-positive associations, overfitting and underfitting, unbalanced datasets, features multicollinearity, and model result interpretability [96]. A major limitation of application of radiomics and ML in head and neck cancer is the small training datasets and the differences in the sizes of the training and

test datasets [95]. The patient sample size can cause incorrect model fitting and, hence, make a model ungeneralizable to new data. Moreover, unbalanced data could be considerably challenging, even in representative cohorts, and might result in unrealistically high model's performance metrics.

Training in ML can be subdivided in two major subcategories: supervised learning, where machines are trained using labelled training data and unsupervised learning, where data are unlabelled, and machines make use of the intrinsic relationship within the data for the purpose of clustering these data. A systematic review that investigated the use of ML models in head and neck cancer radiotherapy, including 48 studies in total, reported the application of numerous machine learning algorithms as presented in Figure 16 [95].

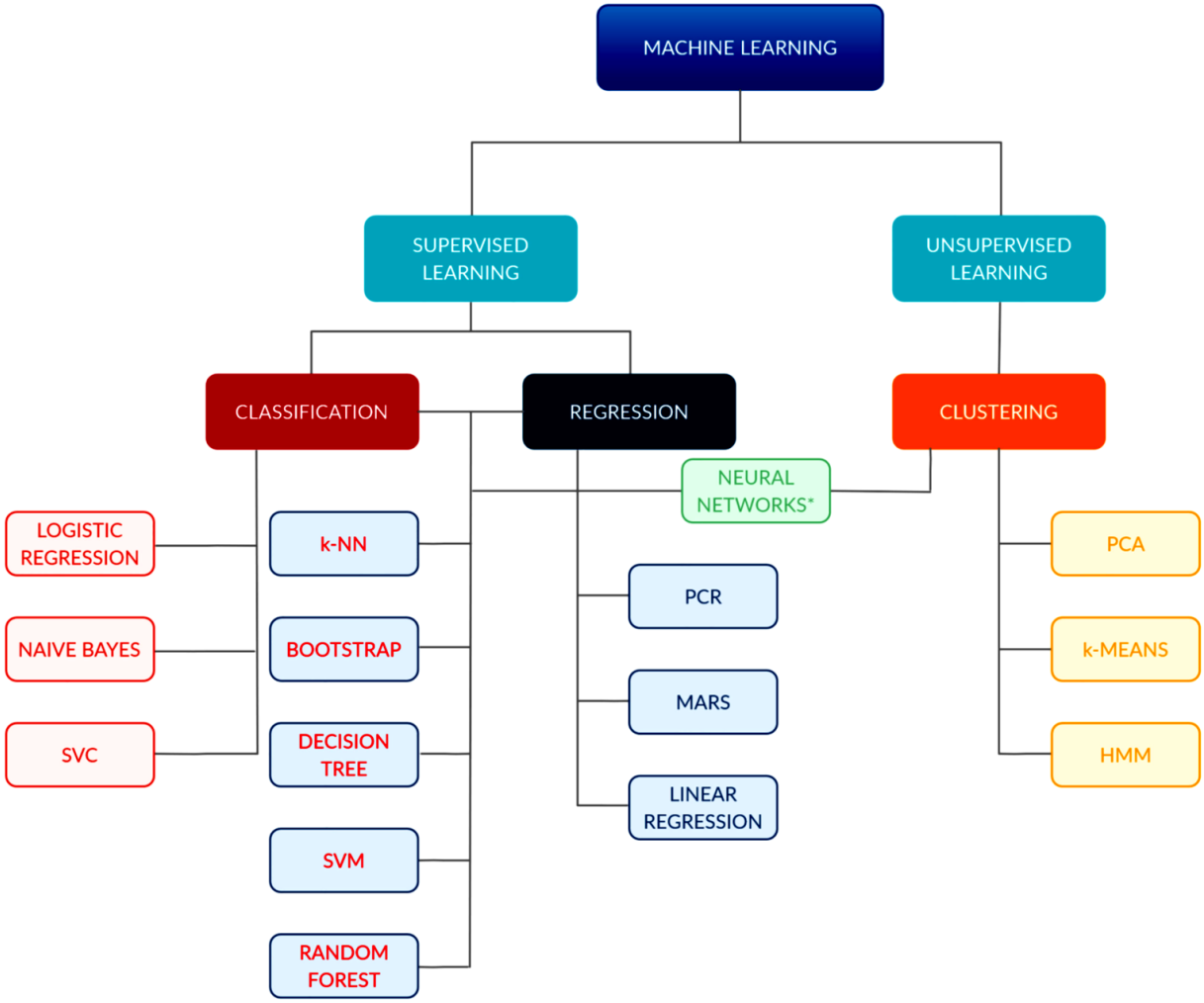


Figure 16: Classification of the machine-learning algorithms included in the analysis. ANN, Artificial Neural Network; CNN, Convolutional Neural Network; FCNN, Fully CNN; HMM, Hidden Markov Model; k-NN, k-Nearest Neighbour; MARS, Multiadaptive Regression Splines; PCA, principal component analysis; PCR, principal component regression; SVC, support vector classifier; SVM, support vector machine.

There are many challenges to overcome before radiomics and machine learning methods become integrated into everyday clinical practice. Collaboration between institutions is essential both for the significant augmentation of the data available, and for the standardization of protocols used for validating the models. Notwithstanding, in the new era of precise medicine, the introduction of new machine learning-derived biomarkers, able to provide significant prognostic power, seems a safe bet [95].

Nonetheless, one should be critical when assessing studies regarding big biomedical data analysis. Due to the inherent complexity of machine learning methods, and the flexibility in specifying these models, results are often insufficiently reported in research articles. Therefore, reliable assessment of those models' validity and consistency can become quite hard. Luo et. al generated a set of guidelines to enable correct application of machine learning models and consistent reporting of model specifications and results in biomedical research [97]. Interestingly, Volpe et al. [95] in their systematic review regarding machine learning for head and neck cancer used an adapted version of the qualitative checklist originally developed by Luo et al. for the quality assessment of the included studies. The organization of the checklist was maintained with the following subsections being rated for each study: "Title and abstract", "Introduction", "Methods", "Results", and "Discussion" allowing for a maximum achievable global score of 58 in their modified Luo classification (Table 8). Results regarding their quality assessment of the included studies are shown in Figure 17 and Figure 18 [95].

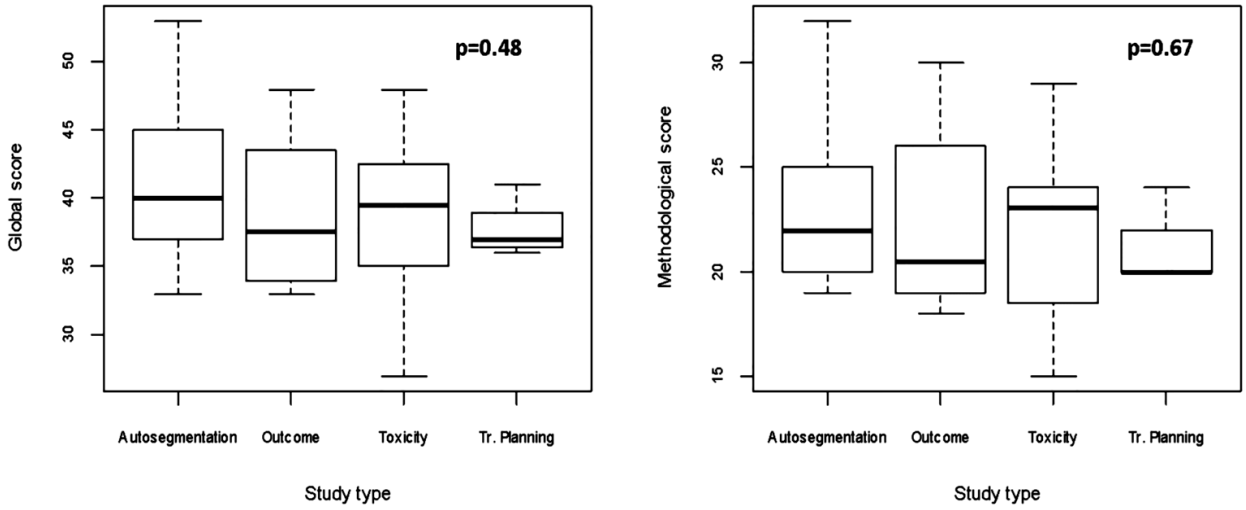


Figure 17: Boxplots for global and methodological scores (modified Luo classification) for the studies included in Volpe et al. systematic review, categorized according to the task of the proposed algorithm(s); Autosegmentation, Outcome, Toxicity, Treatment (Tr.) Planning

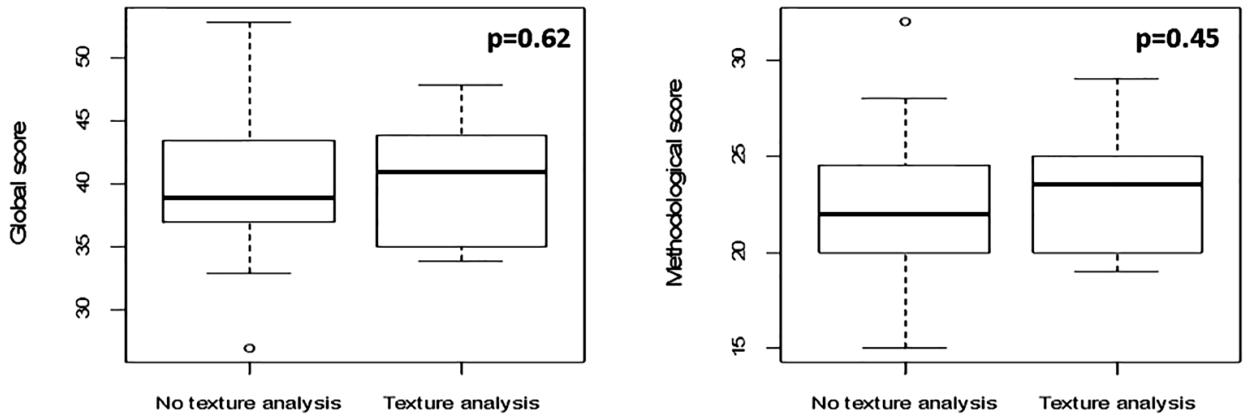


Figure 18: Boxplots representing global and methodological scores (modified Luo classification) for the studies included in Volpe et al. systematic review, categorized per the presence of texture analysis

Table 8: Luo et al. modified quality check list for assessing studies developing and reporting machine learning predictive models in biomedical research as proposed by Volpe et al.

TITLE		
Nature of study	1	Identify the report as introducing a machine learning-based model
ABSTRACT		
Structured summary	2	Background
	3	Objectives
	4	Data sources
	5	Performance metrics of the model or models, in point estimates
	6	Performance metrics of the model or models, in confidence intervals
	7	Conclusion including the practical value of the developed machine learning-based model or models
INTRODUCTION		
Rationale	8	Identify the clinical goal
	9	Review the current practice of any existing models
	10	Review the prediction accuracy of any existing models
Objectives	11	State the nature of study being a machine learning-based model
	12	Define the target of the model
	13	Identify how task resolution may benefit the clinical goal
METHODS		
Describe the setting	14	Identify the clinical setting for the target machine learning-based model
Define the prediction problem	15	Define a measurement for the model task (e.g. patient-based or outcome-based)
	16	Determine that the study is retrospective or prospective
	17	Identify the problem to be prognostic, diagnostic, classification-based, etc.
	18	Determine the form of the model: (1) classification if the target variable is categorical, (2) regression if the target variable is continuous, (3) survival prediction if the target variable is the time to an event
	19	Explain practical costs of prediction errors (e.g. implications of underdiagnosis or overdiagnosis)
	20	Defining quality metrics for the model/models

	21	Define the success criteria (e.g. based on metrics in internal validation or external validation in the context of the clinical problem)
Prepare data for model building	22	Identify relevant data sources
	23	States that relevant data sources were approved by ethics committee or Institutional Review Board
	24	States the inclusion criteria for data
	25	States the exclusion criteria for data
	26	Describe the time span of data
	27	Describe the sample or cohort size
	28	Define the observational units on which the response variable is defined
	29	Define the observational units on which the predictor variable(s) are defined
	30	Define the predictor variables. Extra caution is needed to prevent information leakage from the response variable to predictor variables
Data (feature) pre-processing	31	Describe the data cleaning performed
	32	Describe the transformation performed
	33	Remove outliers with impossible or extreme responses
	34	State any criteria used for outlier removal
	35	State how missing values were handled
Basic statistics of the data set	36	Describe the basic statistics of the dataset, particularly of the response variable
	37	Classification vs. Regression Problem: If classification problem, described ratio of positive to negative classes. If regression problem, describe the distribution of the response variable (e.g. time to event)
	38	Define the model validation strategies: Internal validation: must specify validation strategy (e.g. random split, time-based split, and patient-based split) **(+1 pt),** External validation **(+1 pt)**
	39	Define the validation metrics. (e.g. for regression problems, the normalized root-mean-square error should be used. For classification problems, the metrics should include sensitivity, specificity, positive predictive value, negative predictive value, area under the ROCd curve, and calibration plot)
	40	Retrospective vs. Prospective: For retrospective studies, split the data into a derivation set and a validation set. For prospective studies, define the starting time for validation data collection

Build the predictive model	41	Identify independent variables that predominantly take a single value (eg, being zero 99% of the time)
	42	Report the number of independent variables
	43	Determine a set of candidate modeling techniques (eg. logistic regression, random forest, or deep learning). If only one type of model was used, must also justify the decision for using that model
	44	Define the performance metrics to select the best model
	45	Specify the model selection strategy. (e.g. common methods include K-fold validation or bootstrap to estimate the lost function on a grid of candidate parameter values. For K-fold validation, proper stratification by the response variable is needed)
RESULTS		
Report the final model and performance	46	Report the predictive performance of the final model in terms of the validation metrics specified in the methods section **(+1 pt)** Report the parameter estimates in the model **(+1 pt)** Report the parameter estimates' confidence intervals. **(+1 pt)** When the direct calculation of confidence intervals is not possible, report nonparametric estimates from bootstrap samples
	47	If possible, report what variables were shown to be predictive of the response variable
	48	Designate subpopulation performance characteristics
DISCUSSION		
	49	Interpretation of the final model
	50	Comparison with other models in the literature should be based on confidence intervals
	51	Report the clinical implications derived from the obtained model (e.g. report the dollar amount that could be saved with better prediction. How many patients could benefit from a care model leveraging the model prediction? And to what extent?)
Limitations of the model	52	Sufficient data available for a good fit of the model. In particular, for classification, there should be a sufficient number of observations in both positive and negative classes
	53	Assumed variances in data format: For example, input data format (e.g. inter-scanner variability, sample size, difference in sequences used) or output data format
	54	Potential bias of the data used in modeling
	55	Generalizability of the data

2. MATERIAL AND METHODS

2.1 Material

A Head and Neck PET-CT collection from the cancer imaging archive was retrieved and used for the purposes of this pilot study. The collection was downloaded from the Cancer Imaging Archive [98]. This collection contains FDG-PET/CT and radiotherapy planning CT imaging data of 298 patients from four different institutions in Québec with histologically proven head-and-neck cancer. All patients had pre-treatment FDG-PET/CT scans between April 2006 and November 2014, and within a median of 18 days (range: 6-66) before treatment. These patients were all part of a study described in further detail (treatment, image scanning protocols, etc.) in the publication by [99]. Most of the patients in this cohort (252 patients - 85%) received chemo-radiation with curative intent as part of treatment management. The median follow-up period of all patients was 43 months (range: 6-112).

Three critical factors regarding survival are TNM stage, location of the primary tumour and HPV status. Moreover, sarcopenia has been found as independent adverse prognostic factor for OS and DFS, especially in patients with locally advanced disease, (stage III–IV) [47]. Hence, we considered that it would be beneficial for the purposes of our study to included only patients with Stage IV head and neck cancers, with known site of the primary tumour. We also decided to define premature death separately for each primary site. The minimum follow-up period was set to 5 years. The aforementioned inclusion criteria resulted in 74 patients. Characteristics of the patients both included and excluded, after initial inclusion criteria were applied are shown in Table 9, Table 10, and Table 11.

Table 9: Characteristics of patients excluded due to insufficient follow-up (<5 years)

	Patients excluded because of insufficient follow-up				
Characteristic	**Overall**, N = 123[1]	**Larynx**, N = 4[1]	**Nasopharynx**, N = 4[1]	**Oropharynx**, N = 115[1]	**p-value**[2]
Sex					>0.9
F	31 (25%)	1 (25%)	1 (25%)	29 (25%)	
M	92 (75%)	3 (75%)	3 (75%)	86 (75%)	
Age	62 (57, 70)	72 (67, 75)	63 (60, 68)	62 (56, 69)	0.2
Nodal stage					0.002
N0/1	8 (6.5%)	0 (0%)	3 (75%)	5 (4.3%)	
N2/3	115 (93%)	4 (100%)	1 (25%)	110 (96%)	
Last follow-up (years)	3.37 (2.76, 3.91)	2.88 (1.96, 3.76)	3.33 (3.07, 3.70)	3.37 (2.77, 3.92)	0.6
HPV.status					0.017
-	13 (11%)	1 (25%)	2 (50%)	10 (8.7%)	
+	50 (41%)	0 (0%)	0 (0%)	50 (43%)	
unknown	60 (49%)	3 (75%)	2 (50%)	55 (48%)	

[1]n (%); Median (IQR)

[2]Fisher's exact test; Kruskal-Wallis rank sum test

The vast majority (105/123) of patients excluded due to insufficient follow-up were patients with oropharyngeal carcinomas and HPV (+) or with unknown HPV-status (Table 9). The above two categories of patients could had acted as a confounding factor in our study, given the fact that in patients with oropharyngeal carcinomas survival is more determined by HPV status than by sarcopenia. Therefore, we believe that by excluding these patients we shall not lose important data regarding our clinical problem, and that we will facilitate learning.

In the 74 patients initially included, premature death was defined as death when the survival probability was higher than 75% in the separate, for each primary site, survival curves (Figure 19). Subsequently patients were initially divided in 3 survival categories: (1) patients that died prematurely, (2) patients with overall survival less than 5 years and (3) patients with overall survival more than 5 years.

Table 10: Characteristics of the 74 patients included in the cohort by survival category, after initial inclusion criteria were applied

	Survival category				
Characteristic	**Overall**, N = 74[1]	**<5yr OS**, N = 21[1]	**>5yr OS**, N = 34[1]	**premature death**, N = 19[1]	**p-value**[2]
Primary Site					0.001
Hypopharynx	8 (11%)	6 (29%)	0 (0%)	2 (11%)	
Larynx	15 (20%)	5 (24%)	6 (18%)	4 (21%)	
Nasopharynx	7 (9.5%)	4 (19%)	1 (2.9%)	2 (11%)	
Oropharynx	44 (59%)	6 (29%)	27 (79%)	11 (58%)	
Sex					0.8
F	15 (20%)	3 (14%)	8 (24%)	4 (21%)	
M	59 (80%)	18 (86%)	26 (76%)	15 (79%)	
Age	63 (57, 72)	70 (60, 75)	61 (56, 67)	63 (60, 71)	0.14
Nodal stage					0.5
N0/1	7 (9.5%)	2 (9.5%)	2 (5.9%)	3 (16%)	
N2/3	67 (91%)	19 (90%)	32 (94%)	16 (84%)	
Tumour stage					0.8
T1/2	8 (11%)	1 (4.8%)	5 (15%)	2 (11%)	
T3	21 (28%)	7 (33%)	8 (24%)	6 (32%)	
T4	45 (61%)	13 (62%)	21 (62%)	11 (58%)	

[1]n (%); Median (IQR)

[2]Fisher's exact test; Kruskal-Wallis rank sum test

Further exclusion criteria were applied in the group of patients with oropharyngeal cancers due to the unknown HPV status in the majority of them, which acts as a highly confounding factor. We should also highlight that the exclusion of a potentially confounding group of patients, also resulted in more balanced cohorts in terms of survival categories, which is expected to favour learning.

Table 11: Characteristics of the 74 patients included in the cohort by location of the primary tumour, after initial inclusion criteria were applied

	Location of the primary tumour					
Characteristic	**Overall**, N = 74[1]	**Hypopharynx**, N = 8[1]	**Larynx**, N = 15[1]	**Nasopharynx**, N = 7[1]	**Oropharynx**, N = 44[1]	**p-value**[2]
Sex						0.5
F	15 (20%)	2 (25%)	1 (6.7%)	1 (14%)	11 (25%)	
M	59 (80%)	6 (75%)	14 (93%)	6 (86%)	33 (75%)	
Age	63 (57, 72)	64 (59, 73)	60 (58, 71)	67 (57, 77)	63 (58, 69)	0.8
Nodal stage						0.2
N0/1	7 (9.5%)	0 (0%)	2 (13%)	2 (29%)	3 (6.8%)	
N2/3	67 (91%)	8 (100%)	13 (87%)	5 (71%)	41 (93%)	
Survival						0.001
<5yr OS	21 (28%)	6 (75%)	5 (33%)	4 (57%)	6 (14%)	
>5yr OS	34 (46%)	0 (0%)	6 (40%)	1 (14%)	27 (61%)	
premature death	19 (26%)	2 (25%)	4 (27%)	2 (29%)	11 (25%)	
Time of death (days)	749 (523, 1,096)	615 (409, 1,169)	735 (558, 1,051)	584 (395, 808)	859 (529, 1,116)	0.6
Time of premature death (days)	518 (356, 665)	259 (227, 291)	541 (502, 558)	319 (304, 335)	540 (430, 775)	0.061

[1]n (%); Median (IQR)

[2]Fisher's exact test; Kruskal-Wallis rank sum test

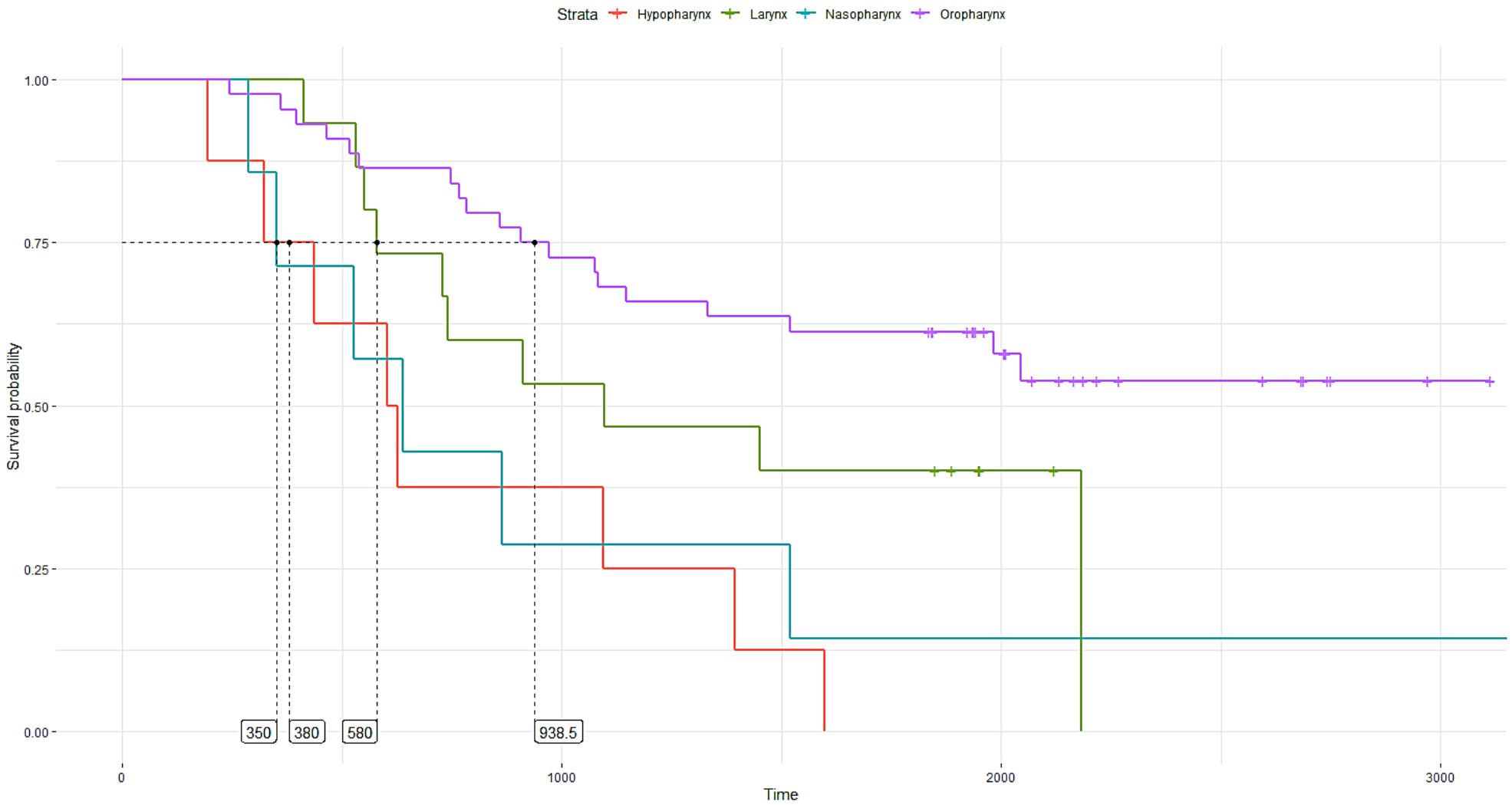


Figure 19: Kaplan-Meier survival analysis by site of the primary tumour. Premature death was defined as death when the survival probability was higher than 75% (lower quartile among those patients who died, separately defined for each primary site)

Two sub-cohorts were created with the application of further exclusion criteria in the group of patients with oropharyngeal carcinomas. One where all patients with oropharyngeal carcinomas and overall survival greater than five years were excluded, resulting in 47 patients (see Table 12), and another one where HPV-negative patients with oropharyngeal carcinomas and overall survival greater than five years were included (meaning 4 extra HPV-negative patients with good prognosis), resulting in a cohort of 51 patients in total (see Table 13).

Table 12: Characteristics of the 47 patients included in the cohort by survival, when all patients with oropharyngeal carcinomas and overall survival > 5 years were excluded

	Survival category				
Characteristic	**Overall**, N = 47[1]	**<5yr OS**, N = 21[1]	**>5yr OS**, N = 7[1]	**premature death**, N = 19[1]	**p-value**[2]
Primary Site					0.015
Hypopharynx	8 (17%)	6 (29%)	0 (0%)	2 (11%)	
Larynx	15 (32%)	5 (24%)	6 (86%)	4 (21%)	
Nasopharynx	7 (15%)	4 (19%)	1 (14%)	2 (11%)	
Oropharynx	17 (36%)	6 (29%)	0 (0%)	11 (58%)	
Sex					0.9
F	8 (17%)	3 (14%)	1 (14%)	4 (21%)	
M	39 (83%)	18 (86%)	6 (86%)	15 (79%)	
Age	63 (59, 73)	70 (60, 75)	59 (58, 66)	63 (60, 71)	0.3
Nodal stage					0.4
N0/1	7 (15%)	2 (9.5%)	2 (29%)	3 (16%)	
N2/3	40 (85%)	19 (90%)	5 (71%)	16 (84%)	
HPV status					>0.9
-	4 (8.5%)	2 (9.5%)	0 (0%)	2 (11%)	
+	4 (8.5%)	2 (9.5%)	0 (0%)	2 (11%)	
unknown	39 (83%)	17 (81%)	7 (100%)	15 (79%)	

[1]n (%); Median (IQR)

[2]Fisher's exact test; Kruskal-Wallis rank sum test

Our final models shall be trained and tested in all three cohorts (the one with 74 patients, the one with 47 patients and the one with 51 patients) and the results will be evaluated accordingly.

Therefore, we will be enabled to test our initial assumption that the smaller cohorts have the potential of favouring training, by suppressing some known confounding factors and by achieving more balanced datasets in terms of survival categories. By having more balanced datasets we will also avoid using oversampling techniques, like Synthetic Minority Oversampling Technique (SMOTE) where synthetic samples are generated for the minority class.

Table 13: Characteristics of the 51 patients included in the cohort by survival, when patients with oropharyngeal carcinomas and overall survival > 5 years were included only if they were HPV(-)

	Survival category				
Characteristic	**Overall**, N = 51[1]	**<5yr OS**, N = 21[1]	**>5yr OS**, N = 11[1]	**premature death**, N = 19[1]	**p-value**[2]
Primary Site					0.2
Hypopharynx	8 (16%)	6 (29%)	0 (0%)	2 (11%)	
Larynx	15 (29%)	5 (24%)	6 (55%)	4 (21%)	
Nasopharynx	7 (14%)	4 (19%)	1 (9.1%)	2 (11%)	
Oropharynx	21 (41%)	6 (29%)	4 (36%)	11 (58%)	
Sex					0.9
F	9 (18%)	3 (14%)	2 (18%)	4 (21%)	
M	42 (82%)	18 (86%)	9 (82%)	15 (79%)	
Age	64 (59, 73)	70 (60, 75)	60 (59, 68)	63 (60, 71)	0.4
Nodal stage					0.7
N0/1	7 (14%)	2 (9.5%)	2 (18%)	3 (16%)	
N2/3	44 (86%)	19 (90%)	9 (82%)	16 (84%)	
HPV status					0.3
-	8 (16%)	2 (9.5%)	4 (36%)	2 (11%)	
+	4 (7.8%)	2 (9.5%)	0 (0%)	2 (11%)	
unknown	39 (76%)	17 (81%)	7 (64%)	15 (79%)	

[1]n (%); Median (IQR)

[2]Fisher's exact test; Kruskal-Wallis rank sum test

2.2 Methods

2.2.1 Statistical analysis

All analyses were conducted using R [100] version 4.0.5 (2021-03-31). For survival analyses, "survival" [101] (version 3.2.11) and "survminer" [102] (version 0.4.9) packages were used. For Bland Altman plots and linear regression models we used the "blandr" [103] (version 0.5.1) package. For assessing the intraclass correlation coefficient (ICC) and the Cohen's kappa coefficient we used the "psych" (version 2.3.9) package [104] and for assessing various evaluation metrics we used the "Metrics" [105] (version 0.1.4) package. The statistical test used in each case is provided along with the results. Statistical significance was defined as a $p < 0.05$.

2.2.2 Selection and pre-processing of CT images, and radiomic features extraction

A dictionary of 4 consecutive CT images at the level of the third cervical vertebra (C3) was created for all 74 patients. These images were manually selected by an

otolaryngology specialist (PM). For the extraction of the radiomic features a flowchart like the one shown in Figure 20 [42] was followed. Image processing usually starts with reconstructed images, which may be processed through several optional steps (e.g., conversion to standardized uptake values, image denoising and image interpolation). Then the region of interest (ROI) can be created automatically, manually or an existing ROI can be retrieved. The ROI is then interpolated as well, and intensity and morphologic masks are created as copies. Radiomics features are then computed from the image masked by the ROI, including features from the intensity histogram (IH), the intensity-volume histogram (IVH), the gray-level co-occurrence matrix (GLCM), the gray-level run-length matrix (GLRLM), the gray-level size-zone matrix (GLSZM), the gray-level distance-zone matrix (GLDZM), the neighbourhood gray-tone difference matrix (NGTDM), and the neighbouring gray-level dependence matrix (NGLDM) families.

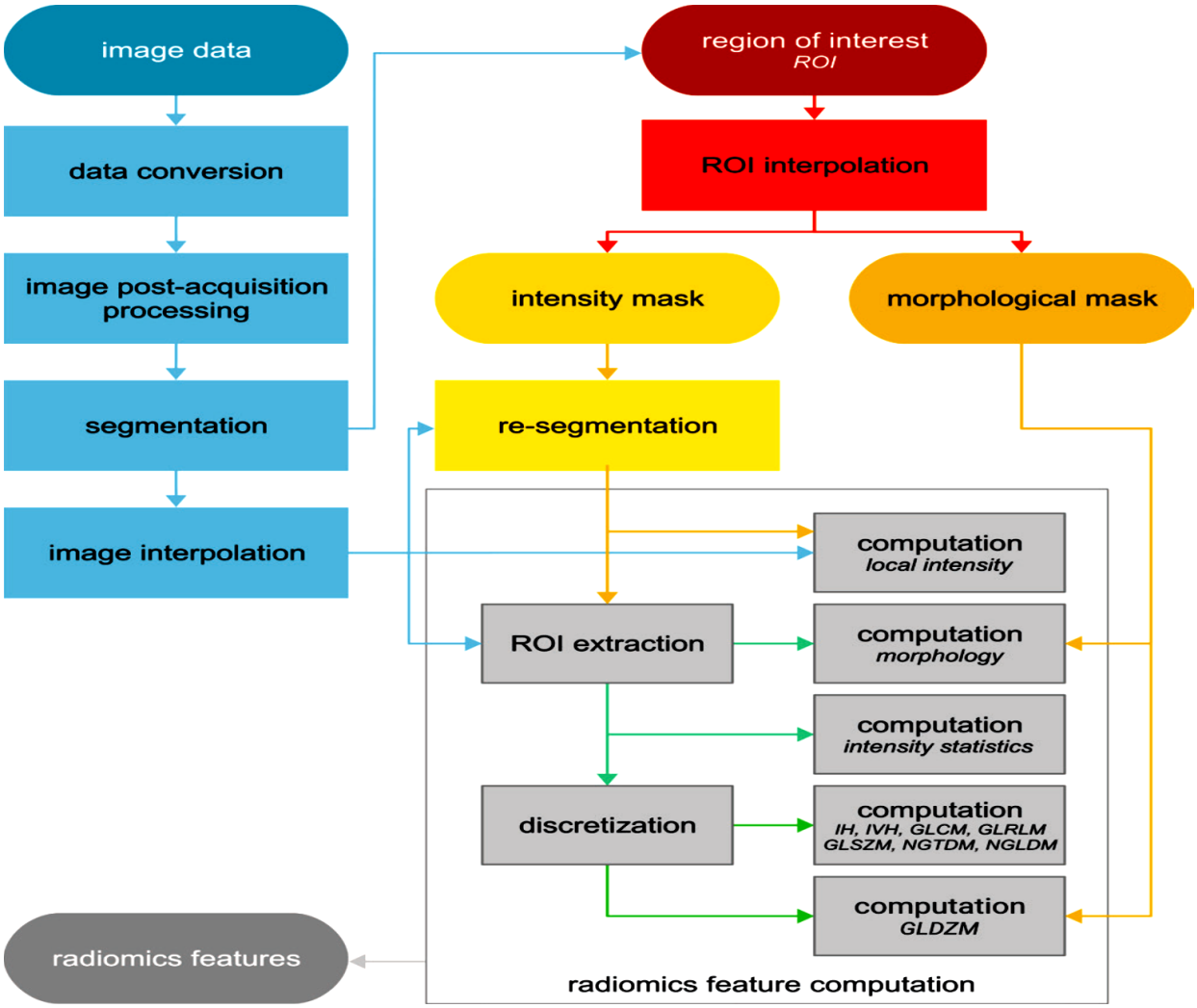


Figure 20: Flowchart of the general radiomics image processing scheme for computing radiomics features

Two approaches of segmentation were followed, manual and automatic segmentation. Moreover, in each segmentation approach, two regions of interest (ROIs) were selected for each CT image. One ROI contained only the paravertebral muscle cross sectional area (CSA) and the other ROI also contained, along with the paravertebral muscles, the intermuscular and intramuscular adipose tissue. Manual segmentation was performed by an otolaryngologist (PM), using 3D slicer desktop software, version 4.11.20210226,

(https://www.slicer.org/) [106]. Automatic segmentation is discussed in more details in a following section. Afterwards, the ROIs' radiomic features were computed in Python using the PyRadiomics package (https://pypi.org/project/pyradiomics/) [107]. For each ROI 93 radiomic features were computed (3 shape related, 15 First order statistics, 24 GLCM, 16 GLRLM, 16 GLSZM, 14 GLDM and 5 NGTDM), resulting in 186 (93x2) computed radiomic features. We also added 3 extra features (fat ratio, muscle ratio and fat/muscle ratio) that were computed using the areas in both ROIs, finally resulting in 189 features. These features were defined as: fat ratio = ((area in ROI with muscles and fat -area in ROI with muscles only) / area in ROI with muscles and fat), muscle ratio = (area in ROI with muscles only / area in ROI with muscles and fat) and fat/muscle ratio = ((area in ROI with muscles and fat -area in ROI with muscles only) / area in ROI with muscles only.

2.2.3 Auto-segmentation of ROIs

Auto-segmentation code was developed in Python v3.8.8 using OpenCV [108] and scikit-image [109] libraries. The auto-segmentation process was based on some anatomical landmarks (mainly the third cervical vertebra) and on the typical Hounsfield Unit (HU) intensities of the different tissues [36], [55], [110]. Figure 22 [110] shows expected radiation attenuation values variation across muscle, fat-infiltrated muscle, and fat. The function developed for the auto-segmentation was taking 4 arguments: the medical image in DICOM format (Digital Imaging and Communications in Medicine is the standard for the communication and management of medical imaging information and related data [111]), followed by the HU windows intensities for muscles, adipose tissue, and bones. While applying this function we used the following HU intensities' upper and lower limits: for bones [150HU,1500HU], for muscles [-20HU,135HU], and for adipose tissue [-200HU, -20HU]). We used -20 HU as lower limit for muscles, because our auto-segmentation code performed better with -20 HU than with -30 HU as lower muscles' attenuation value limit. The main steps that are being executed during the auto-segmentation code that we developed are presented in Figure 22. The first step is to isolate the patient from any external signals such as CT scanner's examination bed (this step is based on a HU window that targets patient's muscles, followed by some morphological operations). The second step is to define the upper limit of our ROI (this step is based on targeting the body of the third vertebra). The third step is to define the rightmost and leftmost borders of our ROI (this step is based on the maximum diameter of the third cervical vertebra along with the vertebra's centroid). The fourth step is to get the mask for our first ROI, that is the paravertebral muscle cross sectional area. The fifth step is to get the mask for our second ROI that is the paravertebral muscles along with the intermuscular and intramuscular adipose tissue. The last two steps use morphological operations and "help" muscles' and fat's masks addition and subtractions, along with HU refinements. Examples of auto-segmentation results compared to manual segmentation are shown in the following figures: Figure 23, Figure 24, Figure 25, and Figure 26.

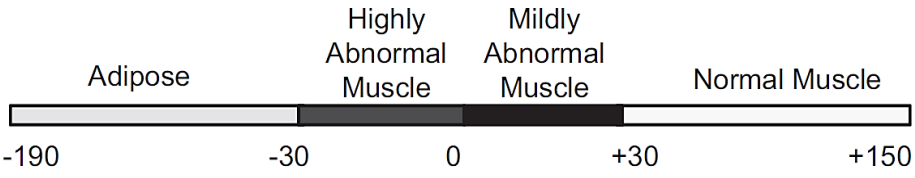


Figure 21: Radiation attenuation values (HU) for muscle, fat-infiltrated muscle, and fat

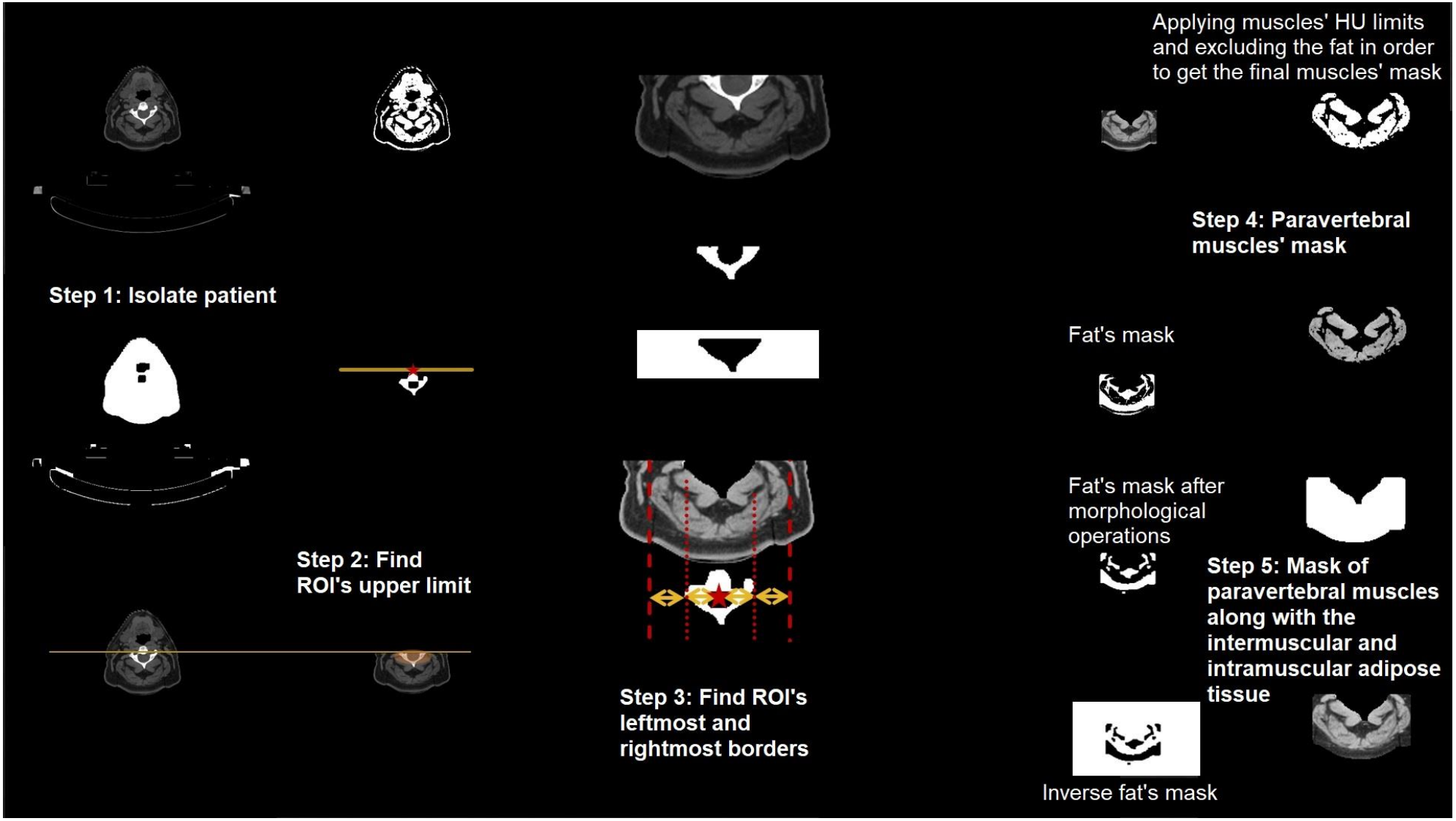


Figure 22: Auto-segmentation step by step

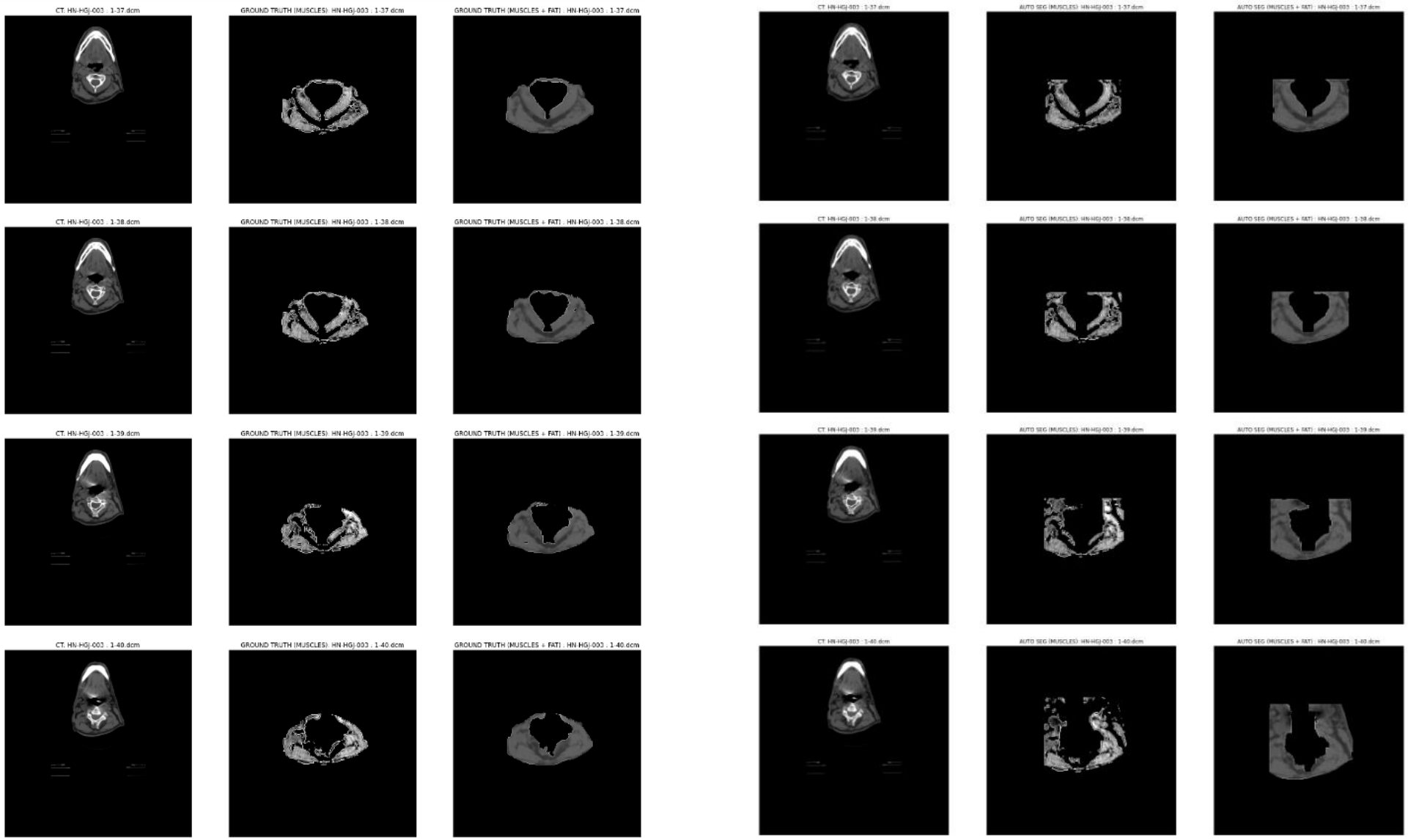


Figure 23: Example of manual segmentation (on the left-hand side) versus auto-segmentation (on the right-hand side) in a male patient with Stage IV laryngeal carcinoma (T3, N2/3) and overall survival < 5 year

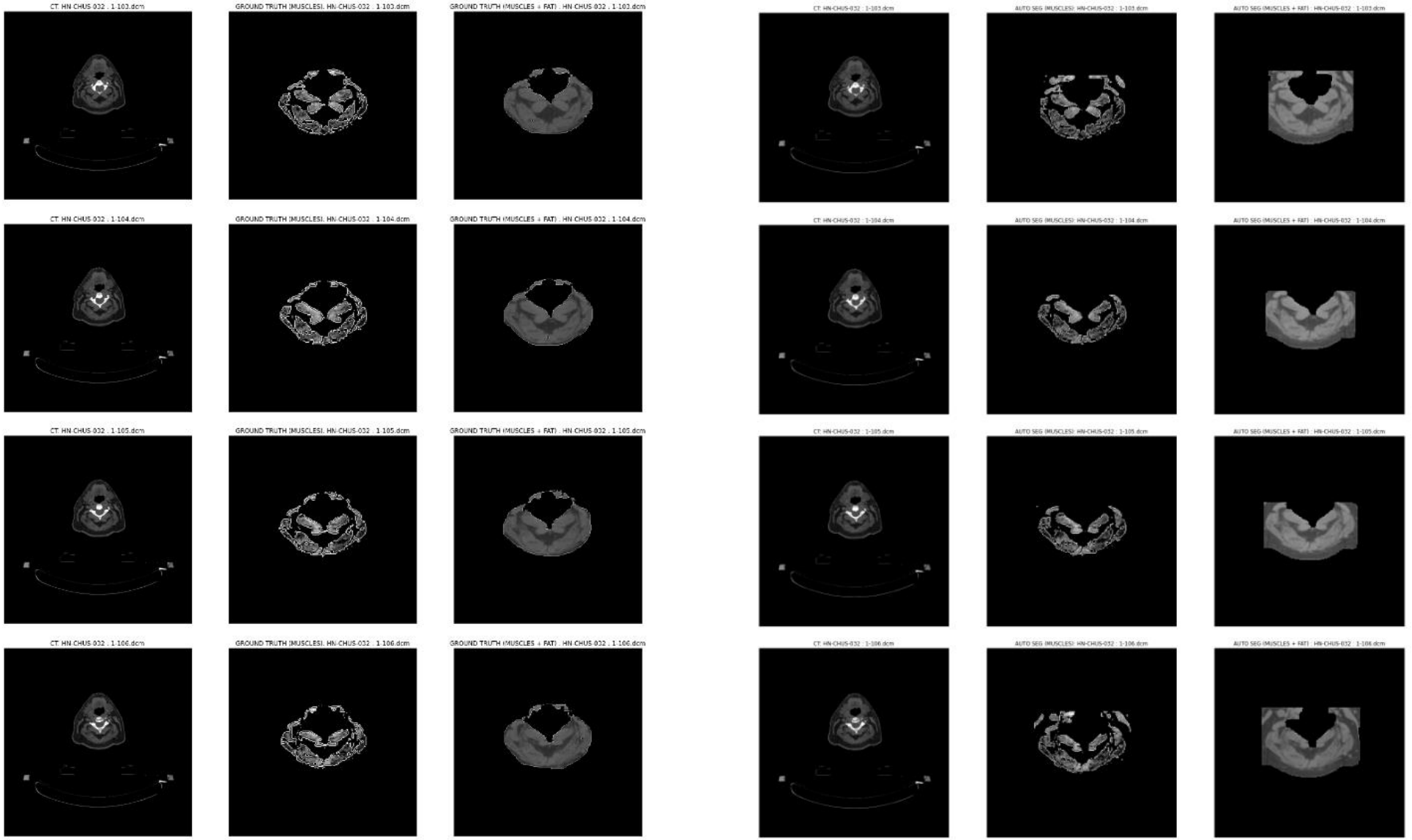


Figure 24: Example of manual segmentation (on the left-hand side) versus auto-segmentation (on the right-hand side) in a male patient with Stage IV laryngeal carcinoma (T3, N2/3) and overall survival > 5 years

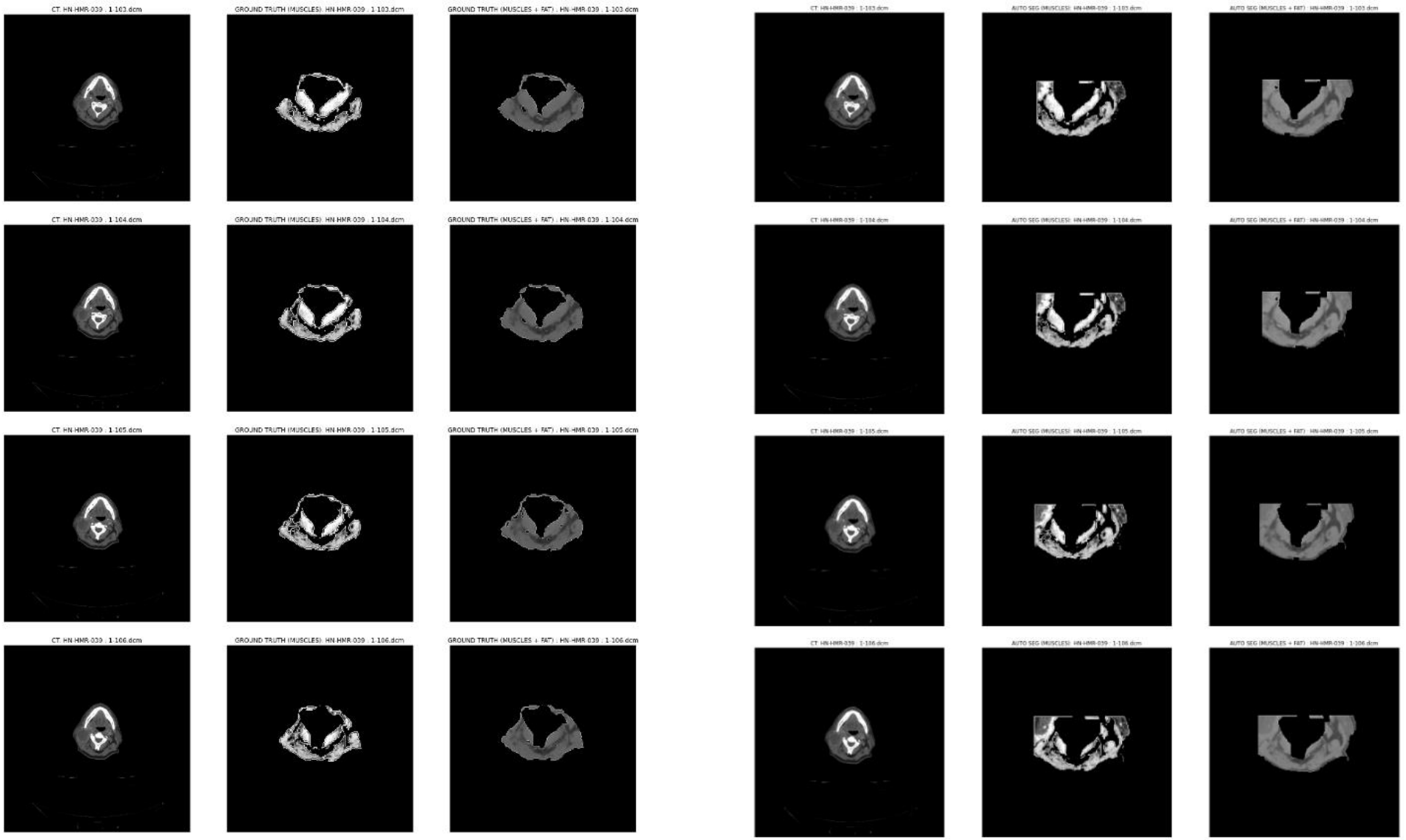


Figure 25: Example of manual segmentation (on the left-hand side) versus auto-segmentation (on the right-hand side) in a male patient with Stage IV carcinoma of the nasopharynx (T1, N2/3), who died prematurely

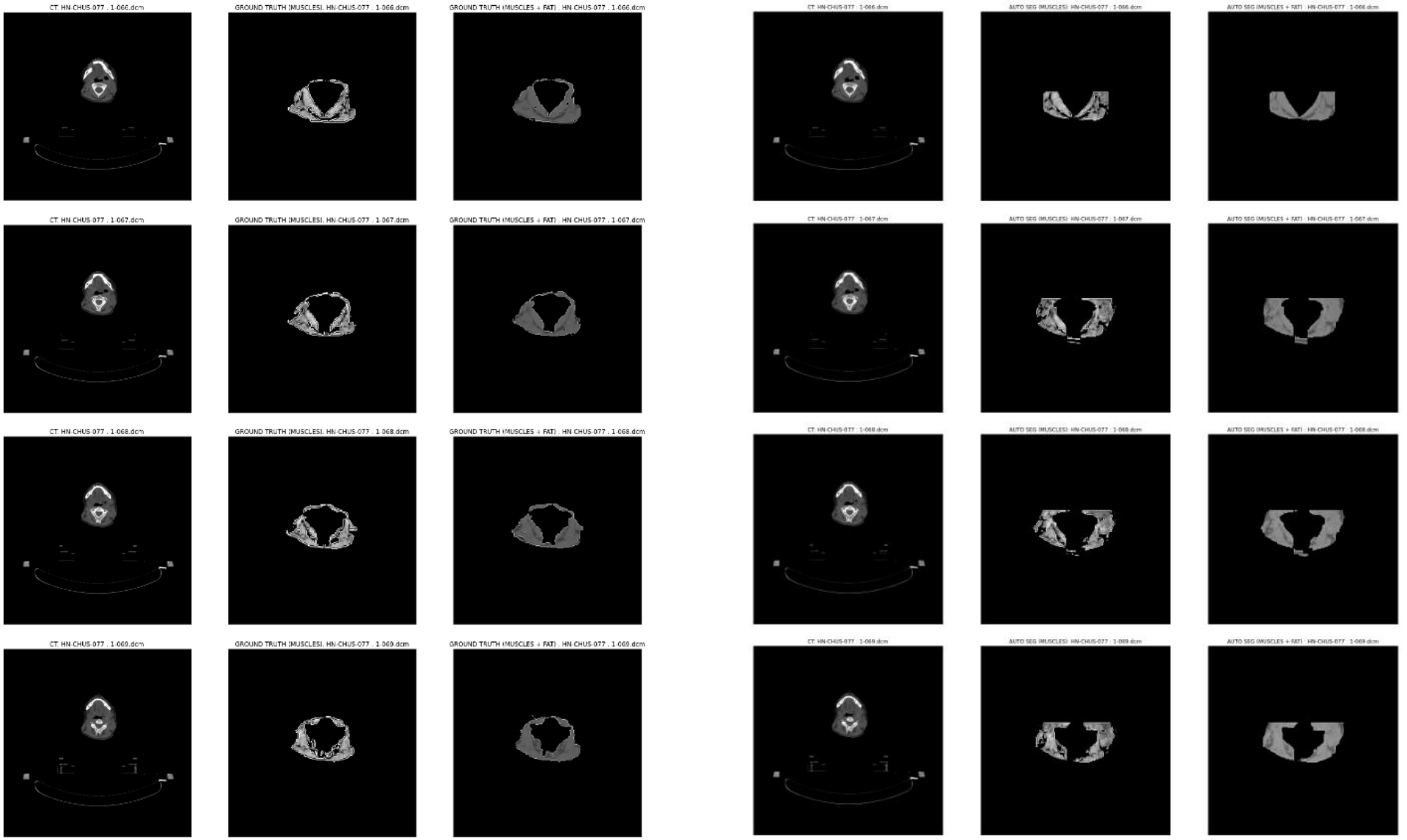


Figure 26: Example of manual segmentation (on the left-hand side) versus auto-segmentation (on the right-hand side) in a female patient with Stage IV HPV-negative oropharyngeal carcinoma (T4a, N2/3), who died prematurely

For comparison purposes, between the manual segmentation method and the automatic segmentation method, we compared the results regarding two features in each one of the two ROIs along with two more features concerning both ROIs (in total 2x2+2= 6 features). We chose to compare features that can be easily clinically interpreted. These features were the mean HU values in each ROI (the one with the paravertebral muscles and the one with the paravertebral muscles along with the intermuscular and intramuscular adipose tissue), the mean HU absolute deviation (distance of all HU intensity values from the mean HU value) in each ROI, the paravertebral muscle ratio, and the adipose tissue / paravertebral muscle ratio. Bland-Altman plots, linear regression models and intraclass correlation coefficient (ICC) for the two methods of segmentation were assessed. The Intraclass correlation is used as a measure of association when studying the reliability of raters, in our case the measurements regarding some features, resulting from the two different segmentation methods (manual and automatic). We used a single-measurement, consistency, 2-way mixed-effects model, to calculate the corresponding ICC results. These results are presented in Table 14. Corresponding Bland-Altman plots along with linear regression model's plots of the two methods' measurements are presented in Figures 27-38. In general, a good degree of agreement was observed between the two methods regarding the examined features. The most prominent difference in the examined features between the two methods was the fact that the auto-segmentation method had a tendency of including a higher proportion of adipose tissue resulting in lower mean HU values, lower muscle ratio and higher adipose tissue / paravertebral muscle ratio. Nevertheless, the linear regression models had very strong correlation coefficients. Altogether, the aforementioned results were quite promising regarding the utility of the auto-segmentation method as a time-efficient alternative approach of extracting computed tomography-derived skeletal muscle related data.

Table 14: ICC results regarding 6 features' measurements resulting from the different segmentation methods (automatic and manual)

Feature	**ICC***	**95% CI**	**Koo & Li (2016) degree of reliability** [112]
Mean HU in paravertebral muscles' ROI	0.929	0.912 < ICC < 0.943	Excellent
Mean HU in ROI with both paravertebral muscles and adipose tissue	0.873	0.843 < ICC < 0.897	Good
Mean HU absolute deviation in paravertebral muscles' ROI	0.848	0.812 < ICC < 0.877	Good
Mean HU absolute deviation in ROI with both paravertebral muscles and adipose tissue	0.717	0.657 < ICC < 0.768	Moderate to Good
Paravertebral muscle ratio	0.853	0.819 < ICC < 0.881	Good
Adipose tissue / paravertebral muscle ratio	0.776	0.727 < ICC < 0.818	Moderate to Good
* 2-way mixed-effects model, single-measurement, consistency			

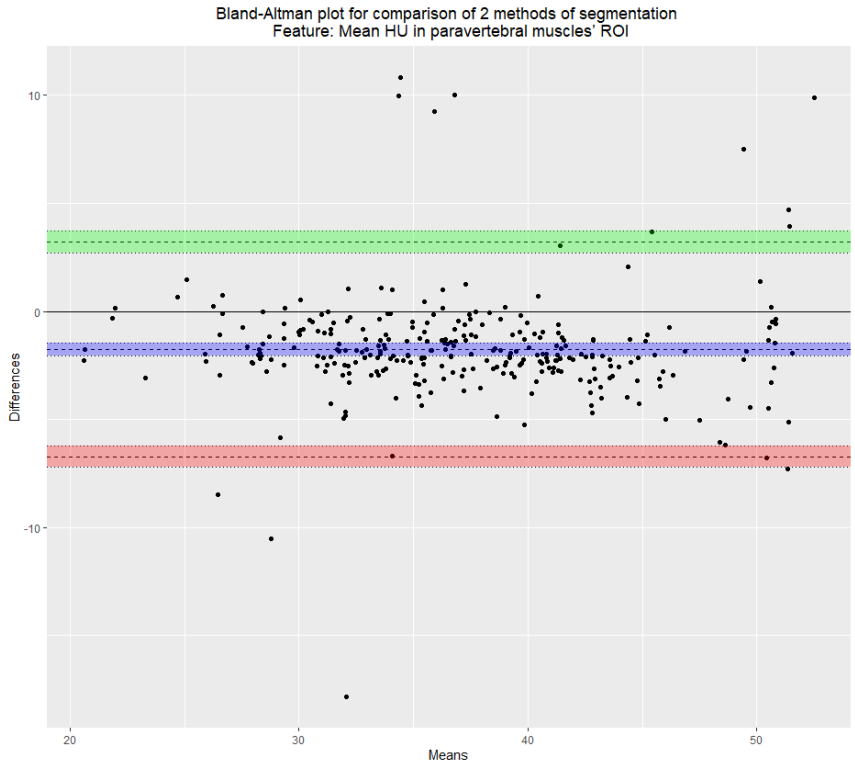
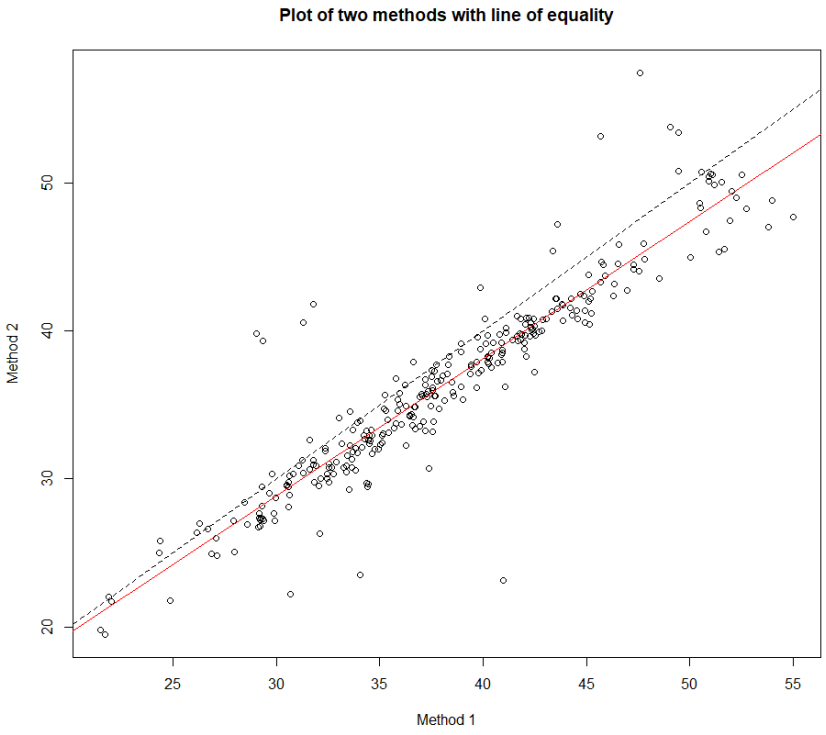


Figure 27: Bland - Altman plot for comparison of auto-segmentation and manual segmentation regarding mean HU in paravertebral muscles' ROI



Method 1: Auto-segmentation, Method 2: Manual segmentation, Correlation coefficient: 0.9293499
Regression equation: Mean HU paravertebral muscles' ROI =
0.9272525 x auto-segmentation's measurements + 1.018074

Figure 28: Plot of the two methods' resulting mean HU in in paravertebral muscles' ROI, with line of equality (dashed black) and linear regression model (solid red)

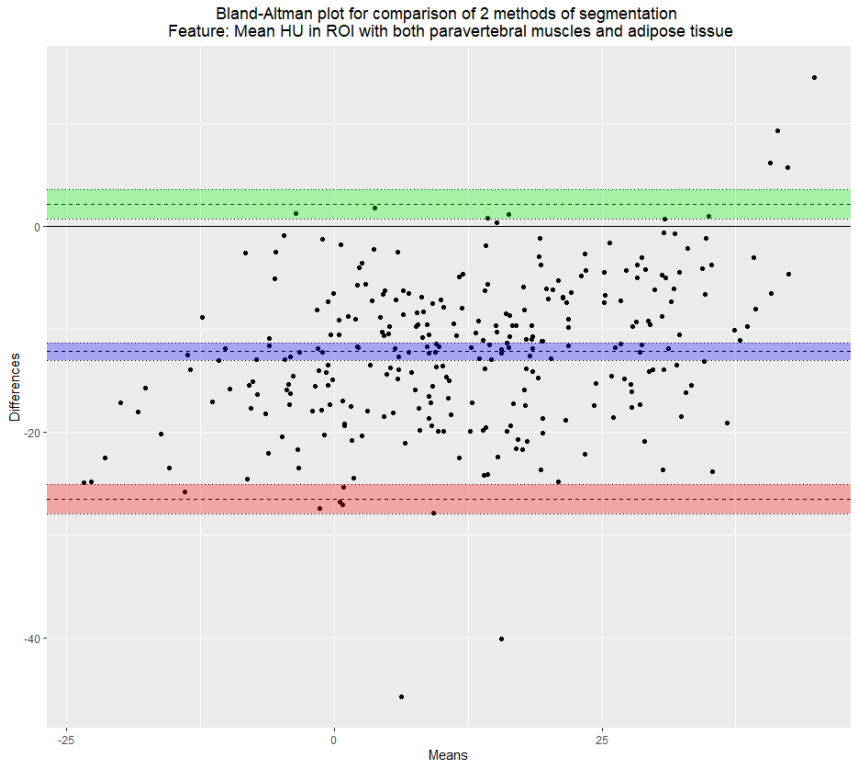
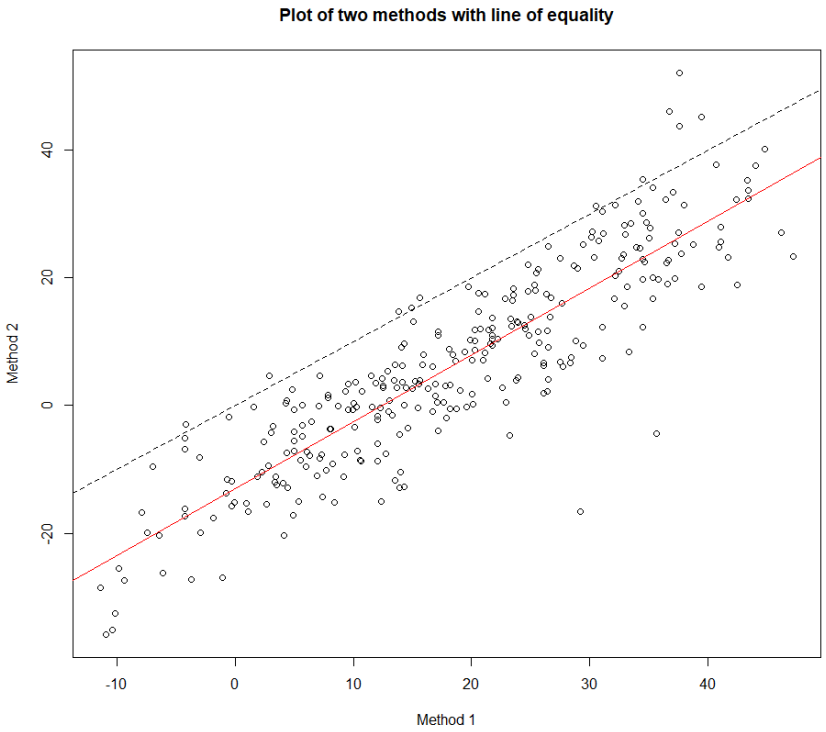


Figure 29: Bland - Altman plot for comparison of auto-segmentation and manual segmentation regarding mean HU in ROI with both paravertebral muscles and adipose tissue



Method 1: Auto-segmentation, Method 2: Manual segmentation, Correlation coefficient: 0.8852279
Regression equation: Mean HU absolute deviation in in ROI with both paravertebral muscles and adipose tissue = 1.04803 x auto-segmentation's measurements -13.05961

Figure 30: Plot of the two methods' resulting mean HU in ROI with both paravertebral muscles and adipose tissue, with line of equality (dashed black) and linear regression model (solid red)

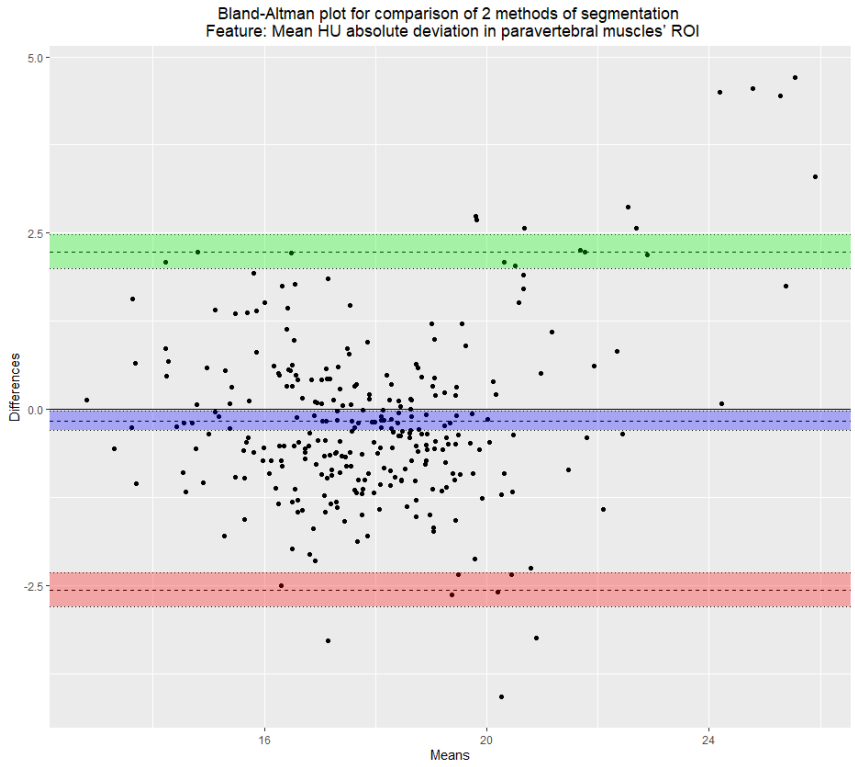
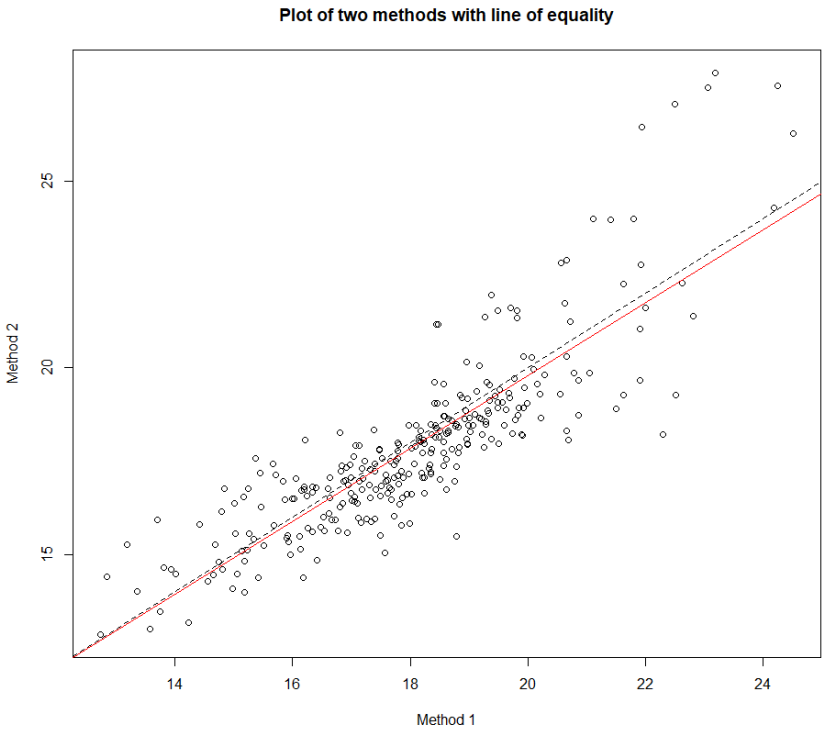


Figure 31: Bland - Altman plot for comparison of auto-segmentation and manual segmentation regarding mean HU absolute deviation in paravertebral muscles' ROI



Method 1: Auto-segmentation, Method 2: Manual segmentation, Correlation coefficient: 0.8549848
Regression equation: Mean HU in paravertebral muscles' ROI =
0.974184x auto-segmentation's measurements + 0.3046604

Figure 32: Plot of the two methods' resulting mean HU absolute deviation in paravertebral muscles' ROI, with line of equality (dashed black) and linear regression model (solid red)

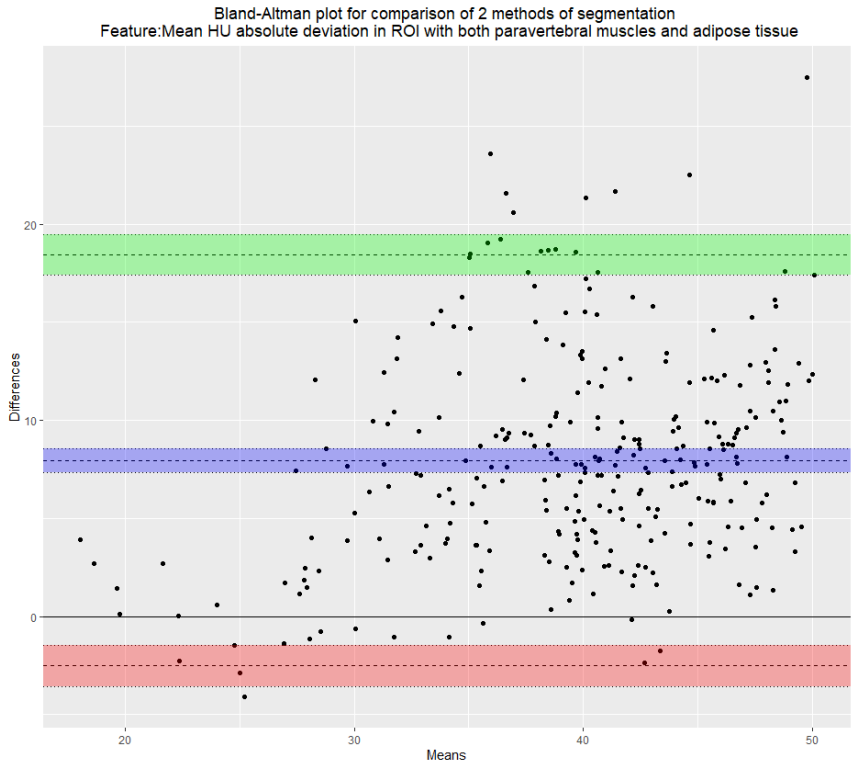
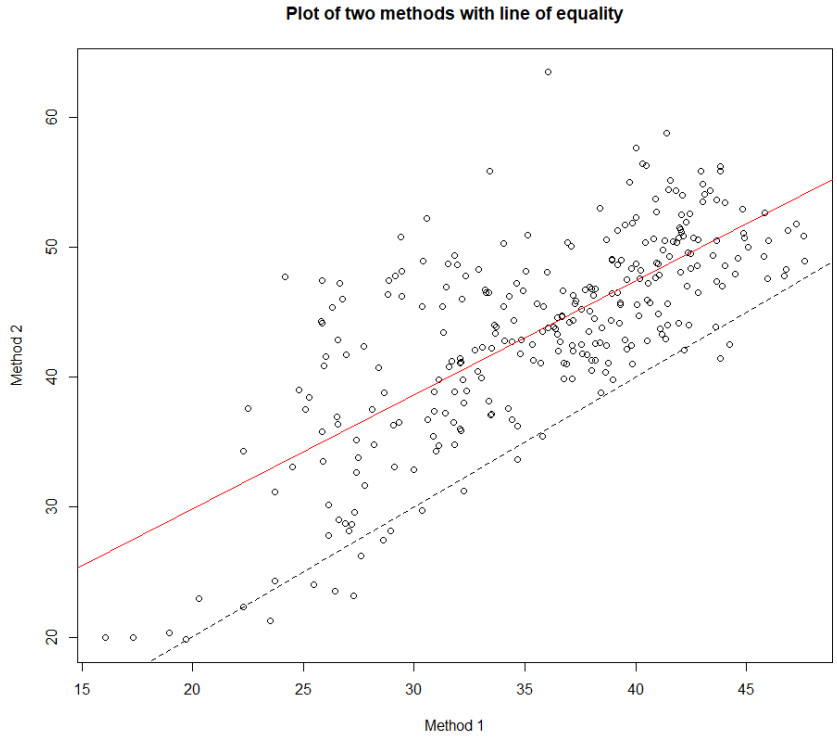


Figure 33: Bland - Altman plot for comparison of auto-segmentation and manual segmentation regarding mean HU absolute deviation in ROI with both paravertebral muscles and adipose tissue



Method 1: Auto-segmentation, Method 2: Manual segmentation, Correlation coefficient: 0.7290292
Regression equation: Mean HU in ROI with both paravertebral muscles and adipose tissue = 0.8763081x auto-segmentation's measurements + 12.37233

Figure 34: Plot of the two methods' resulting mean HU absolute deviation in ROI with both paravertebral muscles and adipose tissue, with line of equality (dashed black) and linear regression model (solid red)

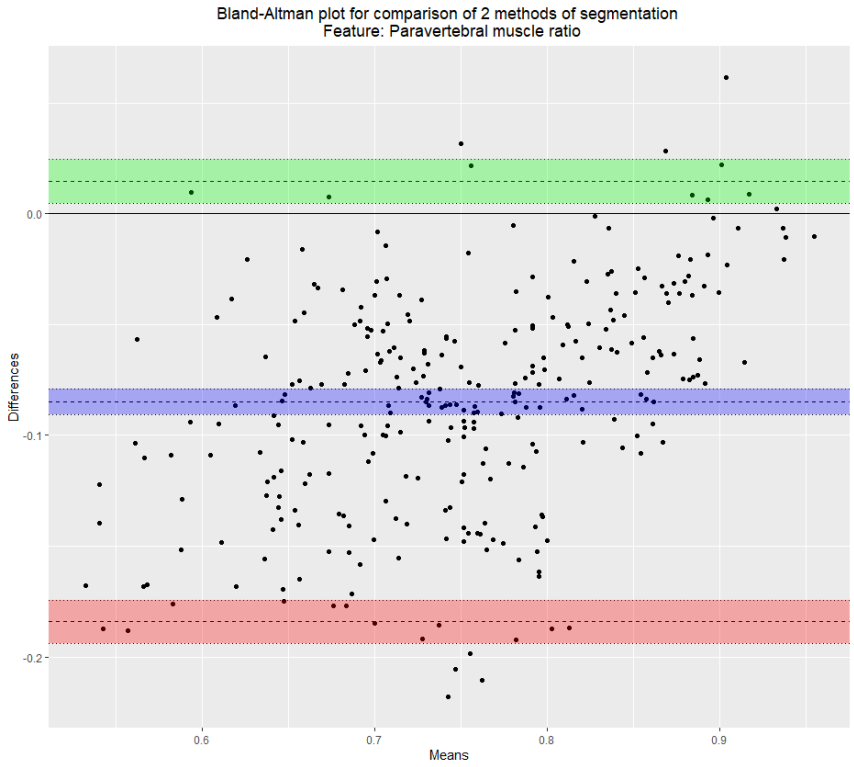
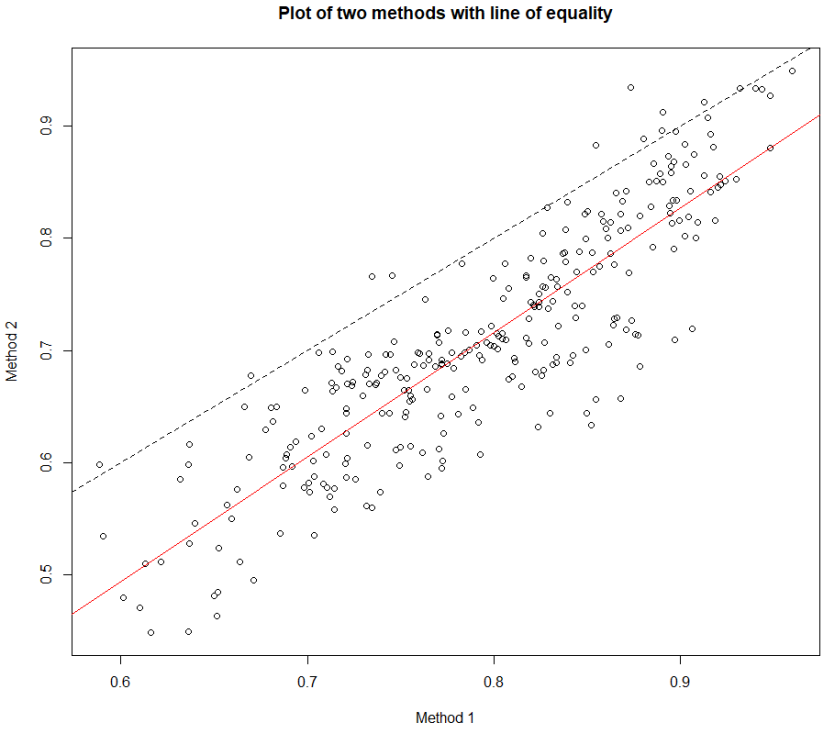


Figure 35: Bland - Altman plot for comparison of auto-segmentation and manual segmentation regarding paravertebral muscle ratio



Method 1: Auto-segmentation, Method 2: Manual segmentation, Correlation coefficient: 0.8768474
Regression equation: Paravertebral muscle ratio =
1.110489 x auto-segmentation's measurements - 0.172471

Figure 36: Plot of the two methods' resulting paravertebral muscle ratio with line of equality (dashed black) and linear regression model (solid red)

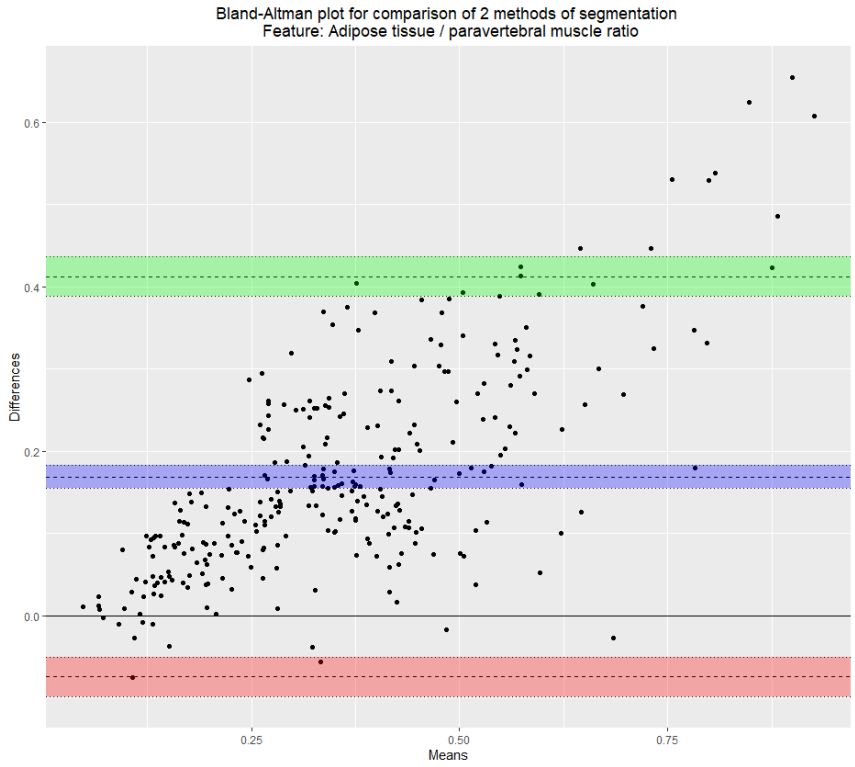
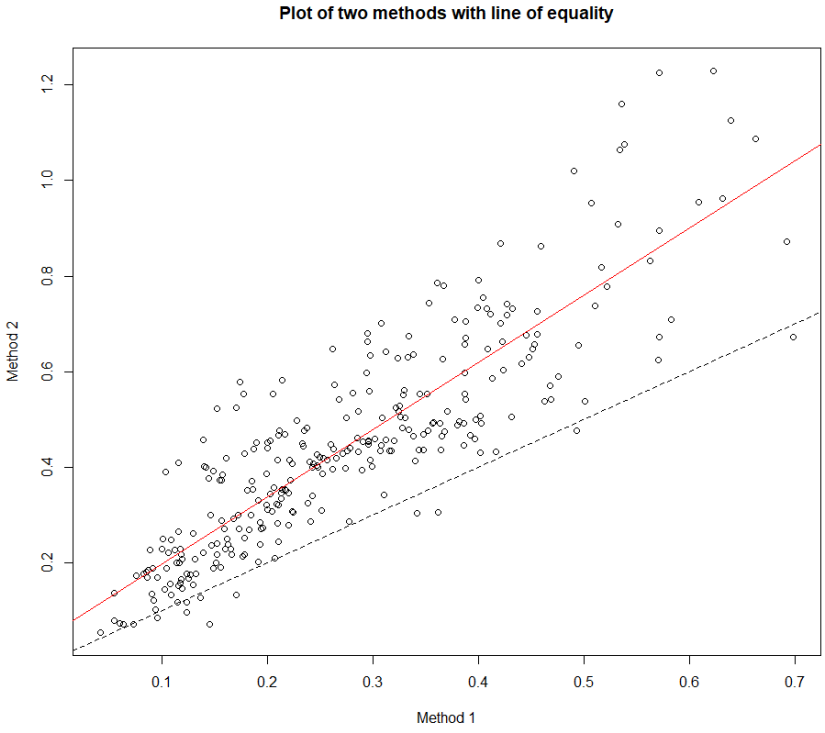


Figure 37: Bland - Altman plot for comparison of auto-segmentation and manual segmentation regarding adipose tissue / paravertebral muscle ratio



Method 1: Auto-segmentation, Method 2: Manual segmentation, Correlation coefficient: 0.8684105
Regression equation: Adipose tissue / paravertebral muscle ratio =
1.407045 x auto-segmentation's measurements + 0.05696887

Figure 38: Plot of the two methods' resulting adipose tissue / paravertebral muscle ratio with line of equality (dashed black) and linear regression model (solid red)

2.2.4 Feature selection

We decided to train our classifiers using only one image per patient, to prevent our classifiers from learning the patient itself and to avoid subsequent overfitting. Therefore, we were left with either 74 or 51 or 47 samples for training, depending on the application of further exclusion criteria in the group of patients with oropharyngeal cancers, while having 189 features per sample. Moreover, in the case of the cohort with the 74 patients, we also had to deal with an unbalanced dataset, regarding our outcome of interest, that is premature death. When dealing with radiomic features, one should be aware that many features could be simply noise, or highly correlated with each other. Hence, feature reduction is necessary to increase prediction accuracy and to minimize computational cost. In general, reducing the number of features can be achieved either supervised or unsupervised. Altogether, when supervised, features are selected based on their discriminative value of outcomes and when unsupervised dimensionally reduction algorithms are being used, maintaining that way more information in the dataset.

A study regarding radiomics-based prognosis analysis for non-small cell lung cancer with 112 patients and 30 radiomic features per patient (11 statistical – first order and 19 textural – second order) [113] addressed the limitations and challenges mentioned above. The authors evaluated the performance of 5 feature reduction techniques (principal component analysis [PCA], independent component analysis [ICA], near zero variance [NZV] , zero variance [ZV], consensus clustering [CC] + PCA) along with no reduction and a filtered feature selection method (Wilcoxon test), using 8 common machine learning classifiers (random forest [RF], generalized linear model [GLM], support vector machine [SVM], naïve Bayes [NB], neural network [NNET], k-nearest neighbour [KNN], mixture discriminant analysis [MDA], partial least squares GLM [PLS]). Moreover, to tackle the problem of unbalanced endpoints in their binary classification problems (death being the most unbalanced outcome with ratio of 0.23) the authors also evaluated the performance of 4 subsampling methods (down sampling, up sampling, Random Over Sampling Examples [ROSE], and Synthetic Minority Over-sampling Technique [SMOTE]). In unbalanced datasets, machine learning algorithms while aiming the highest possible accuracy, have the tendency of sacrificing the minority group, resulting in low sensitivity. Applying subsampling methods, when having more balanced cohorts is not feasible, can significantly improve sensitivity, leading to better predictive performance. Among the feature reduction techniques tested, PCA showed the highest overall (average) predictive performance, RF was the classifier with the highest predictive value and SMOTE was the subsampling method that achieved to enhance AUC (area under the curve) in a balanced way (significantly increasing sensitivity while maintaining high specificity). Thus, the authors proposed the combination of PCA feature selection, RF classifier, and SMOTE subsampling (PCA + RF + SMOTE) as an optimal radiomics pipeline for prognosis of clinical outcomes [113].

Some radiomic features might have less predictive value individually but become important when interaction effects among the features are taken into consideration. Unsupervised feature reduction techniques, like PCA, maintain these interactions favouring the predictive model training process [113]. Moreover, as Aerts et al. [114] demonstrated texture features with higher stability tend to be more informative and have higher prognostic performance as well as reproducibility. Therefore, we critically searched the literature in the last 5 years (2018-2023), regarding human studies (phantom studies excluded) addressing the problem of robustness of radiomic features,

extracted from CT images (using search strings including radiomic feature, robustness, computed tomography +/- perturbation). One relevant review was retrieved, published in 2021. In this review [115] authors found that the most common approach to report the robust features were the percent of robust features, the robust features against all the imaging parameters and the robust feature-parameter that determine which features are robust against which parameter(s). However, authors failed to provide a list of robust features due to the substantial inconsistencies related to the reporting style of the included studies and concluded that radiomics features are dependent on imaging parameters, suggesting that the impact of this dependency must be evaluated on the prediction of clinical outcome.

In a study by X. Teng et al. [116] that used the same Head-Neck-PET-CT collection as we did, but different outcomes and ROI (the region of interest for feature extraction was the primary gross tumour volume) authors assessed radiomic model's reliability using perturbations (authors identified unreliable models by comparing the model's performance on the training dataset with the performance achieved on random perturbations of the training dataset). Aiming to determine whether predictions can be repeatedly produced after perturbations, authors calculated ICC, using one-way model with random effects and absolute agreement, to quantify consistency of the C-index among the samples in the perturbed-train and perturbed-test cohorts. Authors reported a lower training C-index for the perturbed data revealing that evaluating models using their original data is prone to overfitting to noise and to over-estimating the model's learning ability. In their analysis a filter-based feature selection method with two steps (feature–outcome relevance filtering and feature–feature redundancy filtering) was used. To validate the calculation of model robustness, the same experiment was repeated with highly reliable features (ICC > 0.75), leading to a significant increase in the model reliability ICC values from moderate to good, revealing sensitivity of their method to input reliability. In a similar paper X. Teng et al. [117], used an extended dataset, consisted of four publicly available head-and-neck cancer CT collections. Three models were built using all features, good-robust features (ICC > 0.75), and excellent-robbust features (ICC > 0.95). Authors reported that the average model robustness ICC improved significantly from 0.65 to 0.78 ($P < 0.0001$) when using good-robust features and to 0.91 ($P < 0.0001$) when using excellent-robust features. Moreover, by including good-robust features, authors achieved the best average AUC in the unseen data.

Finally, we identified and chose to present, a study [118] that investigated the impact of generative adversarial network (GAN)-based lesion-focused medical image super resolution (SR), on the robustness of radiomic features. Authors applied image SR to increase the number of voxels used since the radiomic features are possibly affected by low statistics in ROI voxels. 75 3D radiomic texture features were calculated (24 GLCM, 14 GLDM, 16 GLRLM, 16 GLSZM and 5 NGTDM). The authors evaluated the robustness of their model's radiomic feature in terms of quantization. Features were extracted from a non-small cell lung cancer CT dataset using different quantization configurations (the number of bins varied [8, 16, 32, 64, 128, 256]). In quantisation of grey levels the number of bins typically has an impact on the GL matrices that are calculated comparing local image intensities, such as co-occurrence (GLCM) and run-length (GLRLM) matrices affecting the values of certain radiomic features. The authors reported that the most important radiomic features in their PCA-based analysis were the most robust features extracted on the GAN-super-resolved images, paving the way for the application of GAN-

based techniques for studies of radiomics for robust biomarker discovery. The highly robust features identified by GAN could possibly generalize well on other CT datasets. The study's results of the robustness analysis related to the textural features (in terms of ICC) according to different image groups resulted in thirteen features that obtained an excellent robustness for at least one of the Original, Cubic and GAN-SR image groups (Table 15 [118]).

Table 15: Features that obtained an excellent robustness for at least one of the image groups (Original, Bicubic, GAN-SR) in a study investigating the impact of GAN-based lesion-focused medical image super-resolution on the robustness of radiomic features

Radiomic feature name	Original	Bicubic	GAN-SR
GLCM Correlation	0.980	0.979	0.984
GLCM DifferenceEntropy	0.846	0.911	0.910
GLCM IDMN	0.996	0.996	0.997
GLCM ID	0.997	0.995	0.998
GLCM MCC	0.633	0.938	0.923
GLCM SumEntropy	0.822	0.897	0.905
GLRLM LongRunLowGrayLevelEmphasis	0.926	0.560	0.631
GLRLM LowGrayLevelRunEmphasis	0.967	0.952	0.944
GLRLM ShortRunLowGrayLevelEmphasis	0.97	0.973	0.925
GLDM DependenceEntropy	0.910	0.870	0.895
GLDM LargeDependenceLowGrayLevelEmphasis	0.985	0.976	0.890
GLDM LowGrayLevelEmphasis	0.986	0.986	0.950
GLDM SmallDependenceLowGrayLevelEmphasis	0.902	0.955	0.946

We decided to test those features with excellent robustness in the manual segmentation cohort with all 74 patients. Having two ROIS (13x2 radiomic features) and adding muscle CSA, fat CSA, fat/muscle ratio and fat ratio we ended up with 30 features per CT image. We then ran multivariate analysis regarding survival (see results in Figure 39).

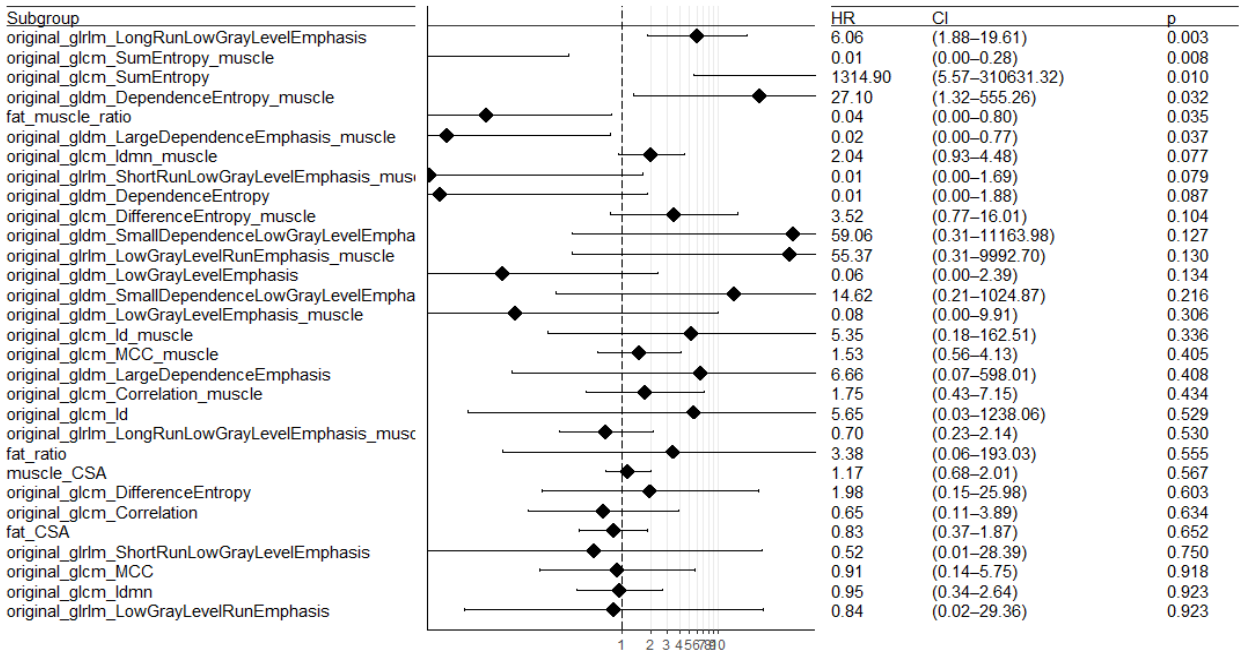


Figure 39: Forest plot based on the results of multivariate analysis of survival (manual segmentation cohort with 74 patients and 4 CT images per patient)

We chose only the 6 features that were found to be significantly related ($p<0.05$) with survival in multivariate analysis, to train our models. Results are presented and discussed in the section "3.1 Results on the manually segmented CT images".

For the auto-segmentation cohort we tested different sets of features using unsupervised clustering. As we will see in the following sections "2.2.5 Unsupervised clustering" and "3.2.1 Unsupervised learning results on auto-segmented CT images", when all features were kept, we achieved the best results. Therefore, based on these results and on the literature [113], we decided to train our models using PCA as feature reduction technique.

2.2.5 Unsupervised clustering

We tested different features sets of the auto-segmentation cohort by applying unsupervised learning algorithms. The sets tested were all features, robust005 (see in Figure 39; "robust" features significantly related [p-value<0.05] with survival in multivariate analysis), robust01 (see in Figure 39; "robust" features with a trend of association [p-value<0.1] with survival in multivariate analysis). We applied Gaussian Mixture Models (GMM) in the 2-dimensional space for better visualization and interpretation of the results. Dimensionality reduction to the 2-dimensional space for each feature set was performed with t-SNE (T-distributed Stochastic Neighbor Embedding) and UMAP (Uniform Manifold Approximation and Projection).

T-SNE [119] takes a set of points in a high-dimensional space and finds a faithful representation of those points in a lower-dimensional space, typically the 2D plane. T-SNE tends to expand dense clusters, contract sparse ones. Notably, distances between well-separated clusters in a t-SNE plot may mean nothing. An important parameter to tune when applying t-SNE is perplexity. Perplexity is kind of a guess about the number of close neighbors each point has. The proposed range is (5:50), and outside this range things get a little weird. With very high values of perplexity, it is highly expected to observe merged clusters whereas with low perplexity values, meaningless "clumping" can occur. In any case, the perplexity should be smaller than the number of points.

On the other hand, UMAP [120] preserves more of the global structure. UMAP's first phase consists of constructing a fuzzy topological representation and its second phase of optimizing the low dimensional representation, to have as close a fuzzy topological representation as possible as measured by cross entropy. UMAP's parameters to tune include min_dist (default 0.1), which refers to the effective minimum distance between embedded points and n_neighbors (default 15) which refers to the size of local neighborhood used for manifold approximation. Smaller values of min_dist will result in a more clumped embedding, where nearby points on the manifold are drawn closer together and larger values of min_dist will result on a more even dispersal of points. Larger values of n_neighbors result in more global views of the manifold, whereas smaller values result in more local data being preserved. In general values of n_neighbors should be in the range (2:100).

A Gaussian mixture model is a probabilistic model that assumes all the data points are generated from a mixture of a finite number of Gaussian distributions with unknown parameters. The expectation-maximization (EM) algorithm is used for fitting mixture of Gaussian models. There are different options to constrain the covariance of the different classes estimated: spherical, diagonal, tied or full. Full means the components may

independently adopt any position and shape, tied means they have the same shape, but the shape may be anything, diagonal means the contour axes are oriented along the coordinate axes, but otherwise the eccentricities may vary between components and spherical is a "diagonal" situation with circular contours. Although one would expect full covariance to have the best performance, it is prone to overfitting, especially on small datasets. GMM were implemented in python using scikit-learn 1.3.2 [121]. GMM's require the user to specify K, the number of Gaussians in the model that are posited to have generated the data. The number of Gaussians corresponds to the number of clusters that the algorithm is looking for. Although GMM is used for clustering, we can also compare the obtained clusters with the actual classes from the dataset.

In the following section "3.2.1 Unsupervised learning results on auto-segmented CT images" we will attempt to evaluate the separability of our dataset for different sets of features and for different number of clusters based on the stratification of our dataset by survival and site of the primary tumor and on the application of further exclusion criteria in patients with oropharyngeal carcinomas; exclusion of patients with OPC who did not die prematurely. Based on the aforementioned, the number of asked clusters will be either 2 (patients with and without premature death; two scenarios with 2 clusters depending on the exclusion or inclusion of patients with OPC who did not die prematurely), or 3 (patients with OPC who died prematurely, non-oropharyngeal carcinoma patients who died prematurely and non-oropharyngeal carcinoma patients who did not die prematurely), or 4 (when no patients were excluded and all patients were stratified by both survival and site of the primary tumor). We will include all 4 images per patient when exploring unsupervised learning's results. Different features set, following dimensionality reduction to 2D space, with either t-SNE or UMAP and with varying algorithms' parameter tuning, shall be tested. Accuracy based on the true clusters (stratified categories of patients) will be documented. Even when searching for more than 2 clusters, we will also calculate binary accuracy depending only on the survival outcome (e.g., if a patient was correctly predicted to die prematurely but incorrectly predicted to belong to a different primary tumor cluster, the prediction will be evaluated as correct in the calculation of binary accuracy).

2.2.6 Supervised learning - training of machine learning classification models

Machine learning classification models were trained using ATOM [122] (Automated Tool for Optimized Modelling) which is an open-source Python package. We kept for training only the 2nd CT slice at the level of C3 per patient. We chose only one image per patient to ensure that our models shall not learn the patient itself. We chose the 2nd CT slice because it usually had the best segmentation quality in the auto-segmentation method. We trained our models using different left-out for test ratios (0.2, 0.25, 0.3, 0.4). We also repeated the training process with 40 different splits of the dataset in order to evaluate our results based on the average metrics' values and on the standard deviation of the metrics' values obtained from the different splits. When using data from the manual segmented images we included all 74 patients, which resulted in a highly imbalanced dataset. In that case we also applied oversampling using the borderline SMOTE algorithm. Figure 40 [123] highlights in an example the differences between algorithms of the SMOTE family. We initially built SVM, RF and XGBoost (eXtreme gradient boosting) classifiers for the manual segmentation case. We also built kSVM (kernel SVM) and RF

classifiers after applying the borderline SMOTE oversampling technique; for distinguishing purposes we will refer to them as SVM_os and RF_os. We then evaluated the results using various ensemble techniques (soft voting of RF and RF_os classifiers with varying weights).

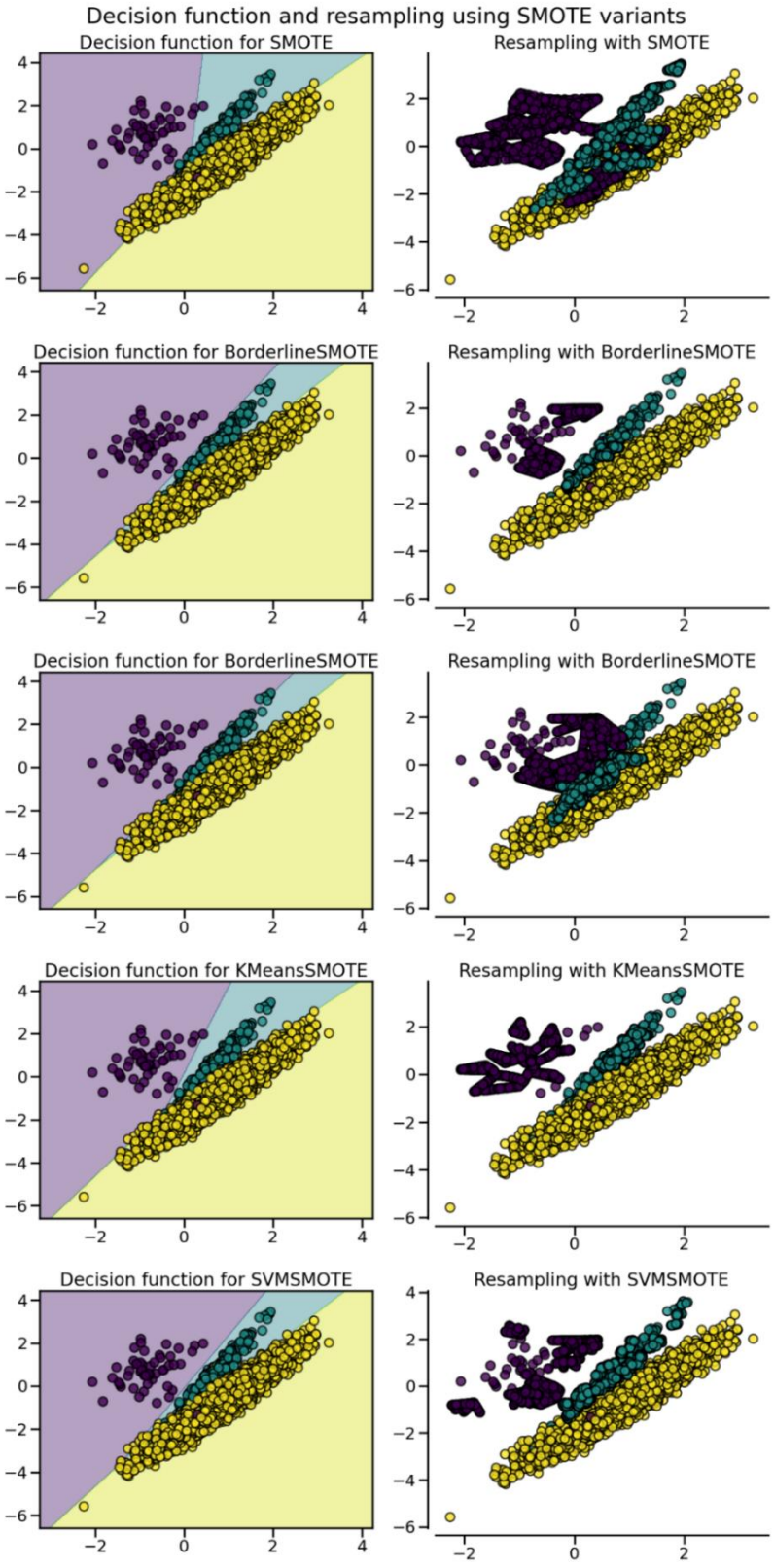


Figure 40: Decision function for different algorithms of the SMOTE family and resulting resampling when used

For the auto-segmentation cohort we built the following classifiers: AdaBoost, GNB (Gaussian Naive Bayes), GP (Gaussian Processes), kSVM, lSVM (linear SVM), MLP (Multi-layer Perceptron), QDA (Quadratic Discriminant Analysis), RF, Trees. We also built soft and hard voting ensemble models by combining with the same weight the vote of multiple classifiers.

Based on the results on the manual segmentation cohort (see section 3.1) and on the fact that HPV status is the main determinant of prognosis in oropharyngeal cancer patients, we decided to exclude oropharyngeal cancer patients with overall survival greater than 5 years. The application of this exclusion criteria resulted in a quite balanced (ratio of premature death: 0.4) sub-cohort of 47 patients (see Table 12). We also trained classifiers in another one sub-cohort with 51 patients (ratio of premature death: 0.37) where we also included HPV (-) OPC patients with OS >5 years (see Table 13). Given that our cohorts had become quite balanced, following the application of exclusion criteria, we did not apply oversampling techniques in those two auto-segmentation sub-cohorts.

2.2.7 Evaluation metrics and validation

The following classification metrics were used to evaluate our results: accuracy, balanced accuracy, precision, recall, Matthews correlation coefficient (MCC), area under receiver operating characteristics curve (AUC-ROC), F1-score, Cohen's kappa coefficient.

Accuracy: correct predictions / number of predictions. It is a metric that should be treated with great caution in imbalanced datasets as it can be quite misleading. In such cases, other evaluation metrics should be considered for better interpretation of the model's utility within the classification problem each time addressed.

Balanced accuracy: ½ (correct positive predictions/number of positives + correct negative predictions/number of negatives). Therefore, by calculating the average accuracy for each class, it can perform better on imbalanced datasets.

Precision: the ratio of true positives and total positives predicted. Precision metric focuses on Type-I errors (false positives). In our case, Type-I error is incorrectly labelling patients who did not die prematurely as high risk for premature death. A precision score towards 1 will signify that our model did not miss any true positives and is able to classify well between high and low risk for premature death.

Recall: the ratio of true positives to all the positives in ground truth. Recall metric focuses on type-II errors (false negatives), in our case, type-II error is incorrectly labelling patients with premature death as low risk for premature death.

Matthews Correlation Coefficient (MCC): MCC is used in machine learning as a measure of the quality of binary classifications and is in essence a correlation coefficient between the observed and predicted binary classifications, returning a value between −1 and +1. A coefficient of +1 represents a perfect prediction, 0 no better than random prediction and −1 indicates total disagreement between prediction and observation.

AUC-ROC; combines the false positive and the true positive rate into a single metric. A receiver operating characteristic curve (ROC curve) is the plot of the true positive against

the false positive rate at each threshold setting and the resulting area under this curve is the AUROC.

F1-score: 2 x (Precision x Recall) / (Precision + Recall). It gives a combined idea about Precision and Recall metrics and is maximum when Precision is equal to Recall. F1-score is considered an effective evaluation metric when true negative is high (our case when all 74 patients are included) and when false positive and false negative are equally costly. F1-score belongs to wider metric family, the F_β-score= $(1+\beta^2)$ x (Precision x Recall) / (β^2 x Precision + Recall). For $\beta=1$ we have F1, for $\beta>1$ we are giving more weight to recall than precision and for $\beta<1$ we are giving more weight to precision than recall. In our case if we wanted to give more weight to one of the two metrics, that would be recall, as our target is to screen for patients in possible high risk for premature death who might be beneficiated by nutritional and other interventions.

Cohen's kappa coefficient: a statistic that is used to measure inter-rater reliability for categorical items that takes into account the possibility of the agreement occurring by chance.

We also used a custom metric, because despite having a binary classification problem (premature death vs no premature death) we wanted to treat with different penalty misclassification of patients with OS > 5 years and of patients with OS < 5 years. In this custom metric, with moderate penalty, scoring was: (+1) for every patient with premature death classified as high risk, (-1) for every patient with OS >5 years classified as high risk, (-0.5) for every patient with premature death classified as low risk, (0) for every other case e.g., patient with OS < 5 years classified either as high or low risk. The sum was then divided by the number of patients who died prematurely. Thus, the maximum score would always be 1, whereas the minimum will be a varying negative.

Moreover, we assessed the robustness of the model trained using the bootstrap technique, which creates several new data sets (we used 20) selecting random samples from the training set (with replacement) and evaluates them on the test set (we used mcc as evaluation metric for bootstrapping).

Finally, when evaluating the models trained with ROIs from CT images obtained with the automatic segmentation method, we used as external validation set the three images per patients, that hadn't been used for training (Figure 41). We averaged the three risks (either hard voting or soft voting derived from probabilities) from the different classifiers to obtain an average risk for premature death for each patient. Then we calculated the optimal risk cut-offs regarding survival separately for each one of the 40 different runs (train-test splits of the dataset). We evaluated classifiers in terms of the standard deviation of the optimal cut-offs obtained. Afterwards, we tested classifiers with the lowest optimal cut-offs' standard deviation. We set cut-off for each classifier's risk as either the median or the mean of all the optimal cut-offs. By setting the same cut-off (not always the optimal) for each one of the 40 scenarios, we tried to be as much unbiased as possible, given that finding the optimal cut-off is an outcome orientated method. We then tested survival outcomes in all 40 scenarios in terms of the resulting p-values from the comparison between the survival curves of the two risk groups of patients (low and high risk for premature death).

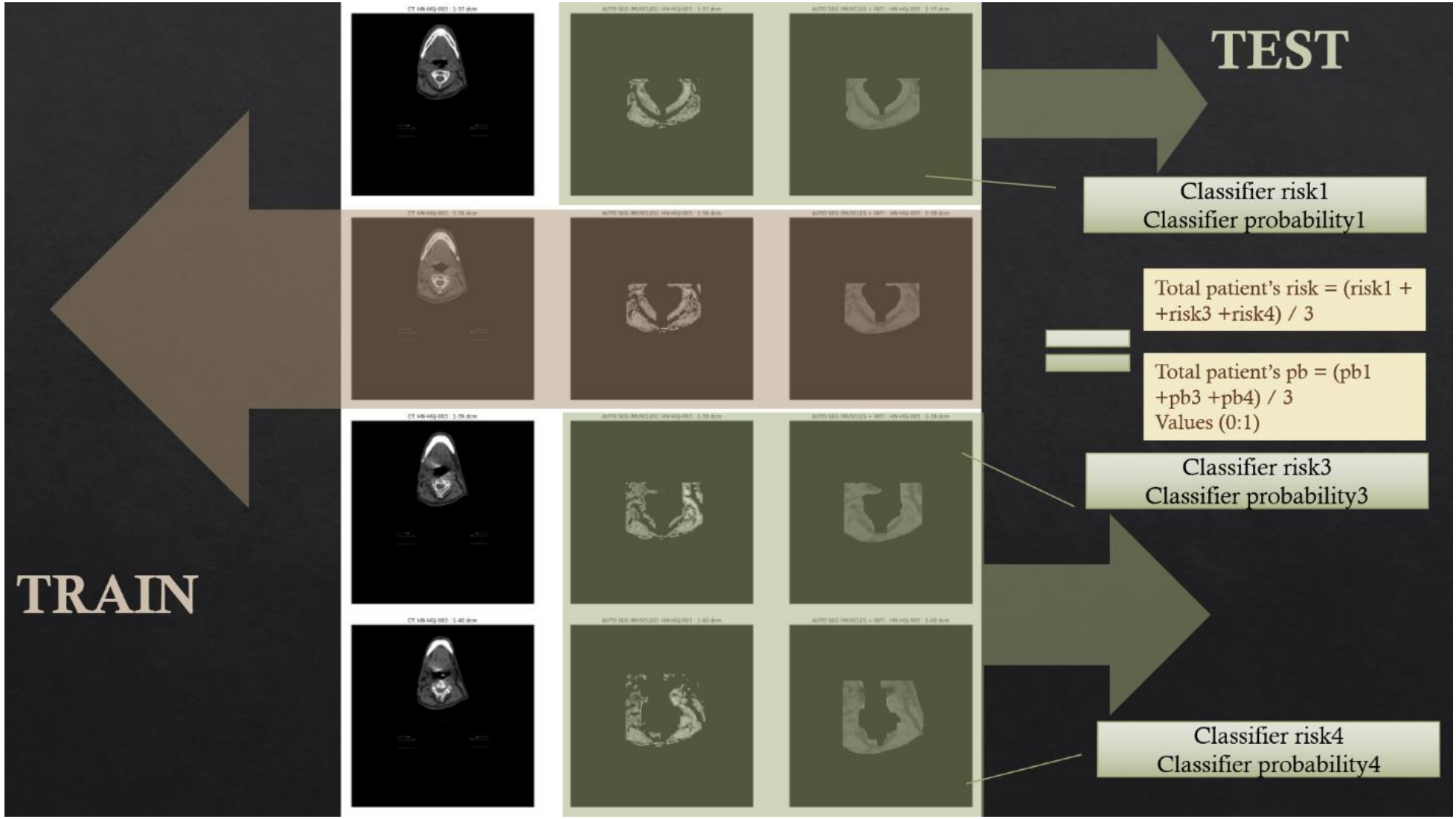


Figure 41: Using as external validation set the three images per patient (1st, 3rd and 4th CT slice at the level of C3) that remained completely unseen when training our classification models (only images from the 2nd CT slice were used for training)

3. RESULTS

3.1 Results on manually segmented CT images

The following classification results derived from training in the manual segmentation cohort (with all 74 patients), using only one image per patient for training (the 2nd CT slice at the C3 level) and as feature set only the 6 "robust" features which were found to be significantly related ($p<0.05$) with survival in multivariate analysis (see section 2.2.4). RF classifiers showed better performance in terms of average balanced accuracy compared to XGBoost and kSVM classifiers (see Figure 42).

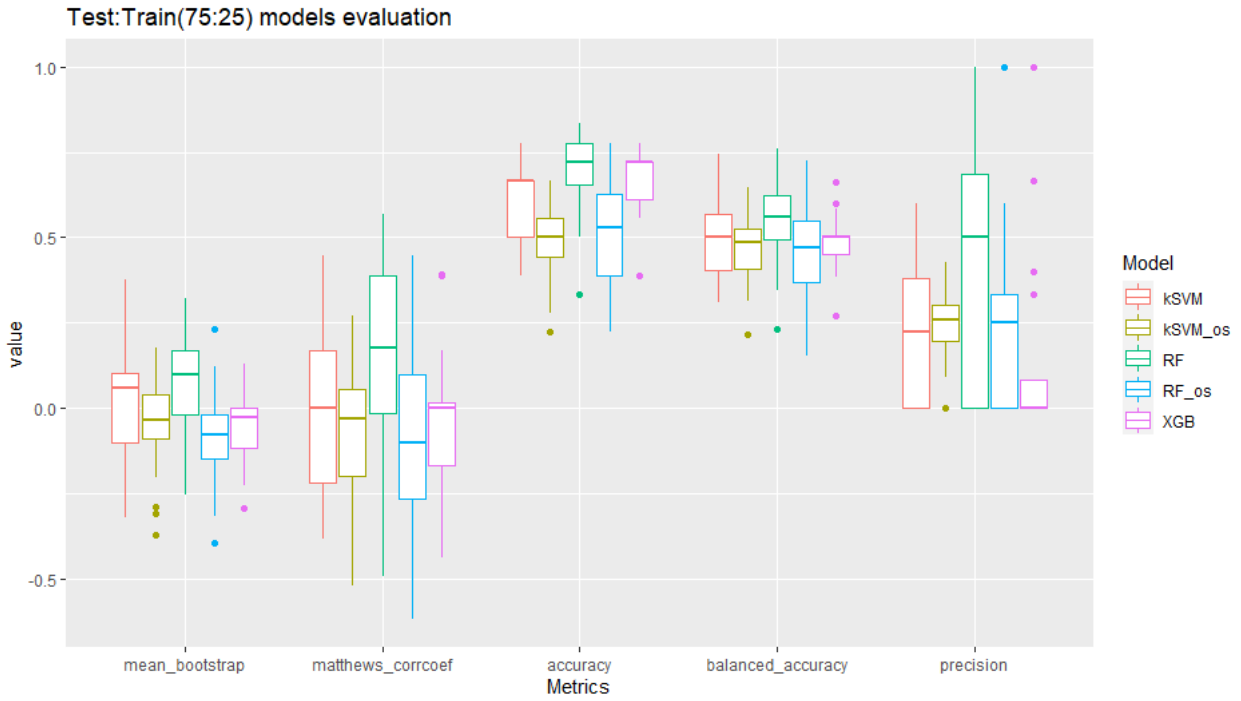


Figure 42: Classifiers' evaluation when trained with a 0.25 ratio left for testing (40 different test:train splits of the dataset) in the manual segmentation cohort

Although oversampling with BorderlineSMOTE did not improve balanced accuracy, we investigated if a combination of RF classifiers trained with and without oversampling would improve the classification results. Therefore, we tested results from soft-voting ensembles were RF and RF_os contributed with different weights (75% RF + 25% RF_os, 70% RF + 30% RF_os, 60%RF + 40% RF_os, 50% RF + 50% RF_os, 40% RF + 60% RF_os). We compared results in terms of the variance of the optimal cut-off for survival stratification (see Figure 43) and in terms of the custom metric (see section 2.2.7 Evaluation metrics and validation), which took also into consideration the subgroups with OS > 5 years and OS < 5years (see Figure 44). In general, the combination of RF with RF_os significantly improved the results achieved by RF and RF_os alone. The best results derived from the 60% RF + 40% RF_os combination. We also investigated the performance of the 60% RF + 40% RF_os combination in different subgroups of patients within the cohort (oropharyngeal carcinomas, non-oropharyngeal carcinomas, HPV(+) oropharyngeal carcinomas and laryngeal, hypopharyngeal and HPV(-) oropharyngeal carcinomas; Figures 45,46,47,48). Finally, we checked the proportion of positive custom metric scores achieved with the 60% RF + 40% RF_os voting scheme (see Figure 49),

and how moderate scores (near 0) or even negative scores are being translated in the survival curves with a few examples (see Figures 50,51).

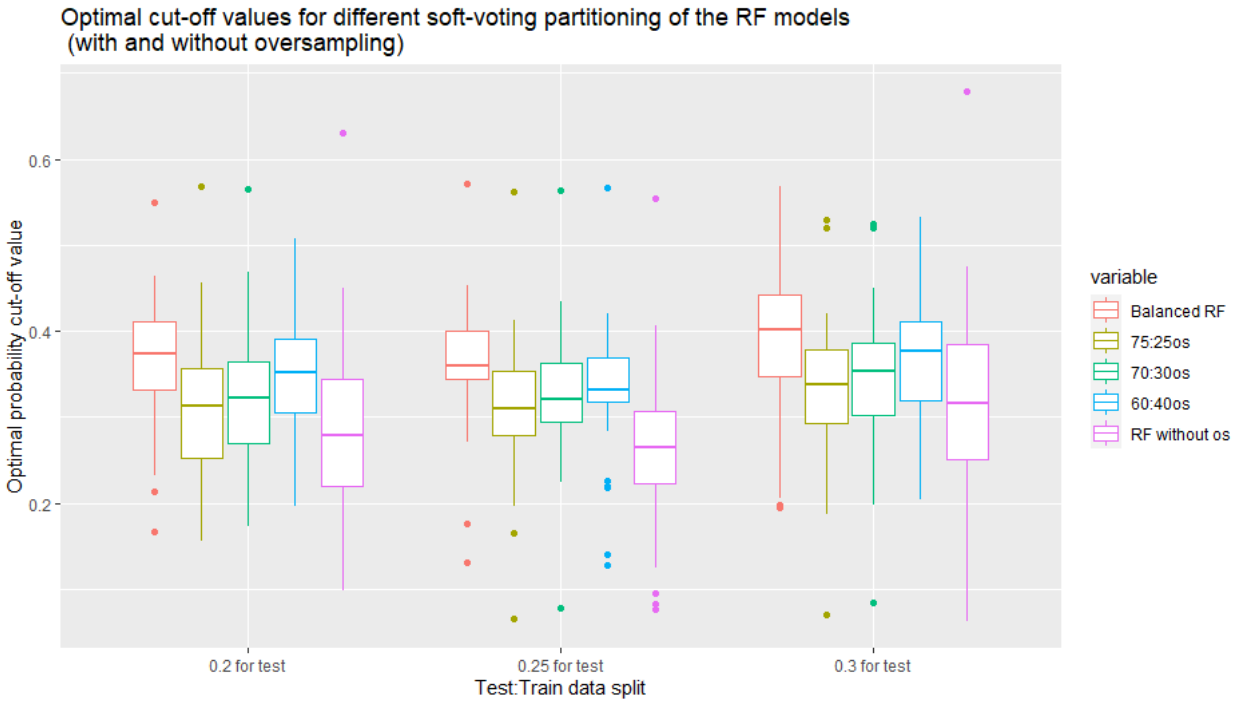


Figure 43: Optimal cut-off values regarding survival stratification for different soft-voting partitioning of the RF models trained with and without application of oversampling techniques

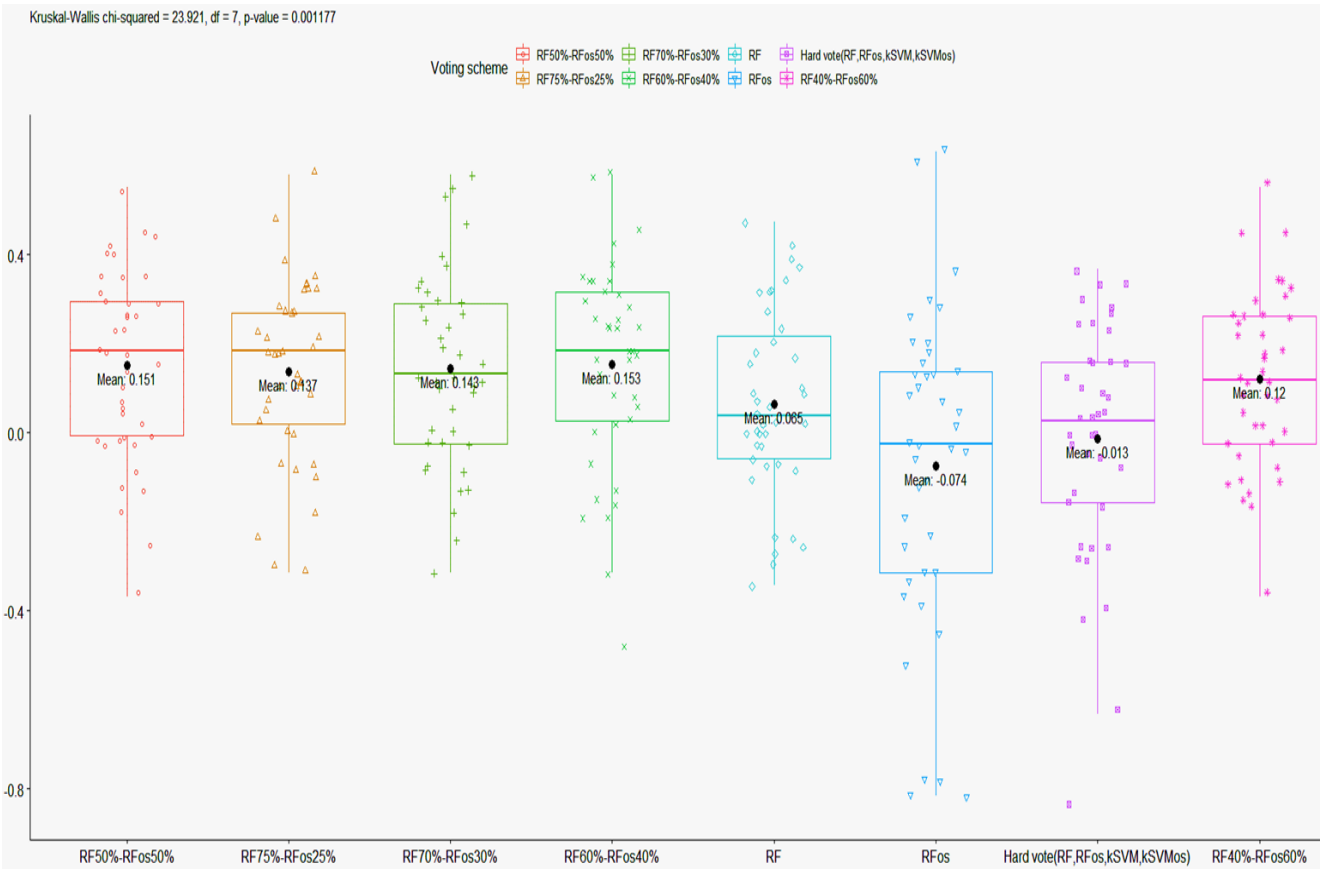


Figure 44: Scoring results for different voting schemes (40 different train:test splits with a 0.25 ratio left for testing)

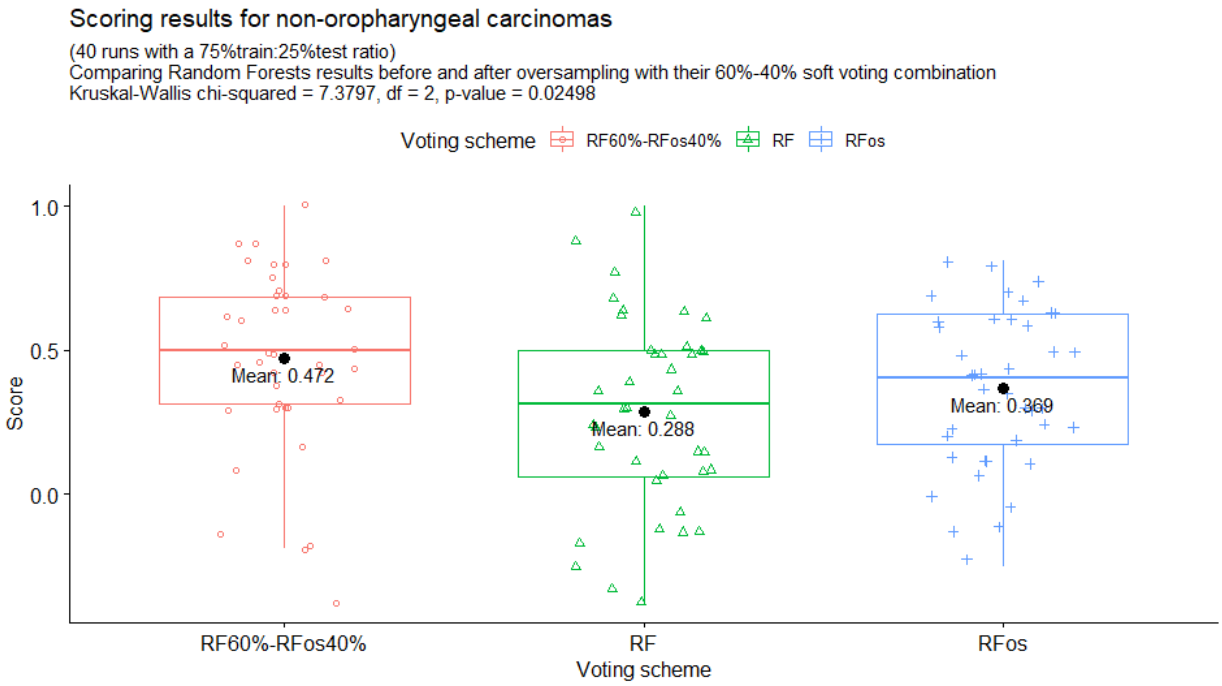


Figure 45: Comparing RF, RF_os and ensemble RF60%-RF_os40% scoring results in manual segmentations' non-OPC sub-cohort

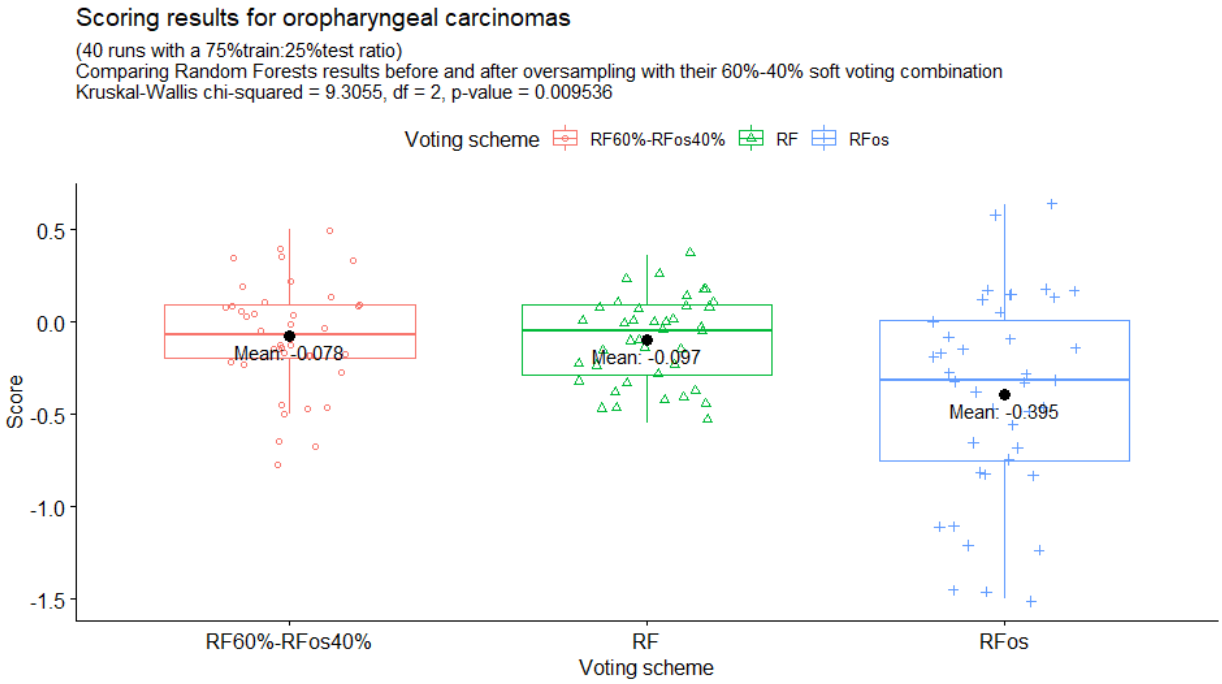


Figure 46: Comparing RF, RF_os and ensemble RF60%-RF_os40% scoring results in manual segmentations' OPC sub-cohort

The soft voting combination RF60%-RF_os40% achieved better results on both subgroups of patients regarding the primary tumour site (oropharyngeal and non-oropharyngeal carcinomas), whereas RF and RF_os classifiers alone performed better in different subgroups. Moreover, overall results were considerably better in the non-OPC subgroup. Interestingly, regardless the disappointing results of the RF_os classifier in the

OPC subgroup when combined with the RF classifier managed to improve the RF classifier's results. Scoring results in HPV(+) patients were poor, something expected as HPV status heavily affects survival and HPV(+) patients have significantly better prognosis regardless of other factors.

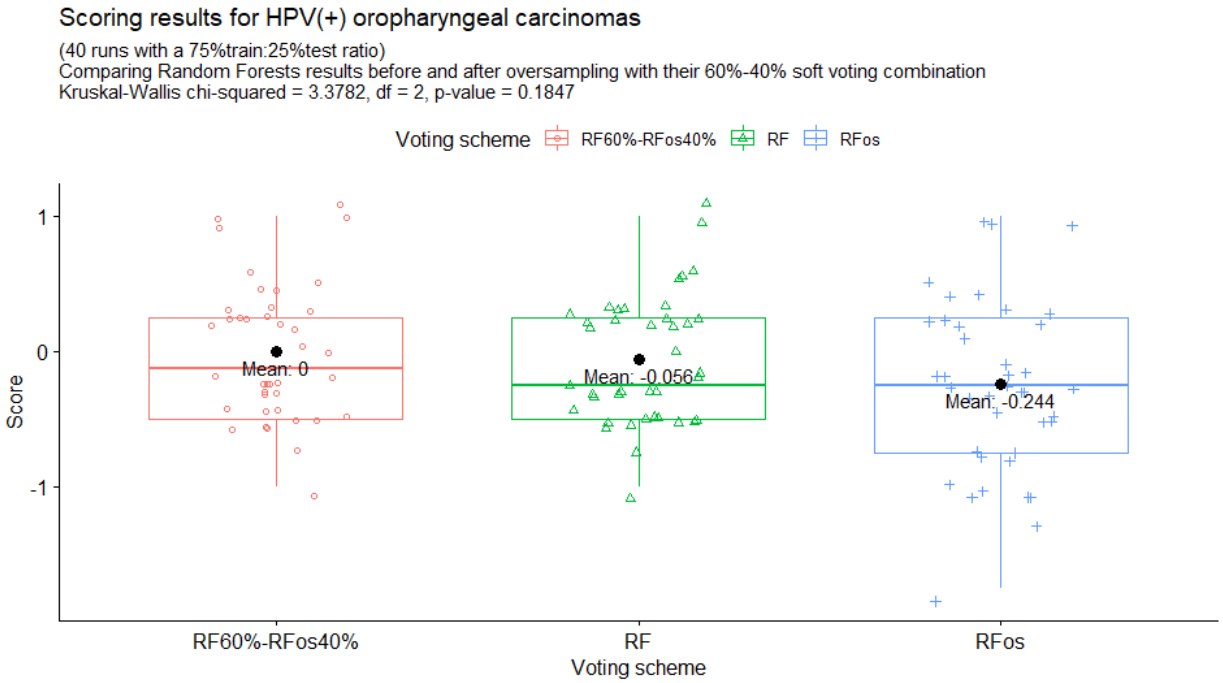


Figure 47: Comparing RF, RF_os and ensemble RF60%-RF_os40% scoring results in manual segmentations' HPV(+) OPC sub-cohort

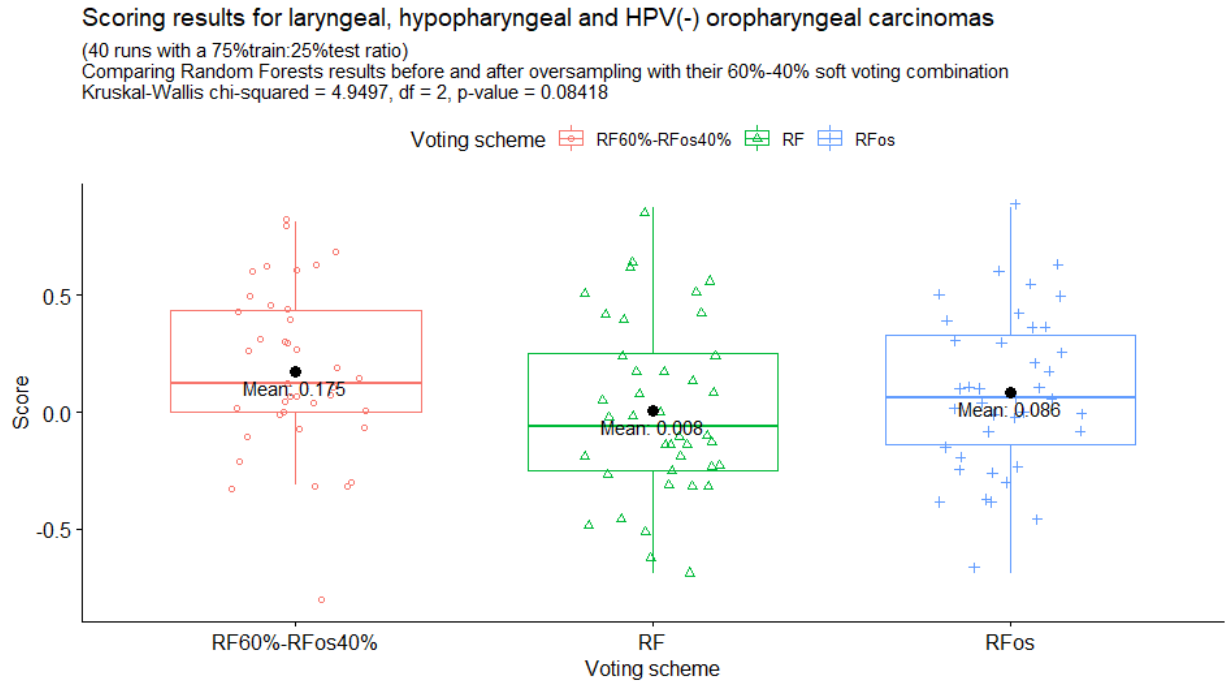


Figure 48: Comparing RF, RF_os and ensemble RF60%-RF_os40% scoring results in manual segmentations' sub-cohort including only laryngeal, hypopharyngeal and HPV(-) oropharyngeal carcinomas

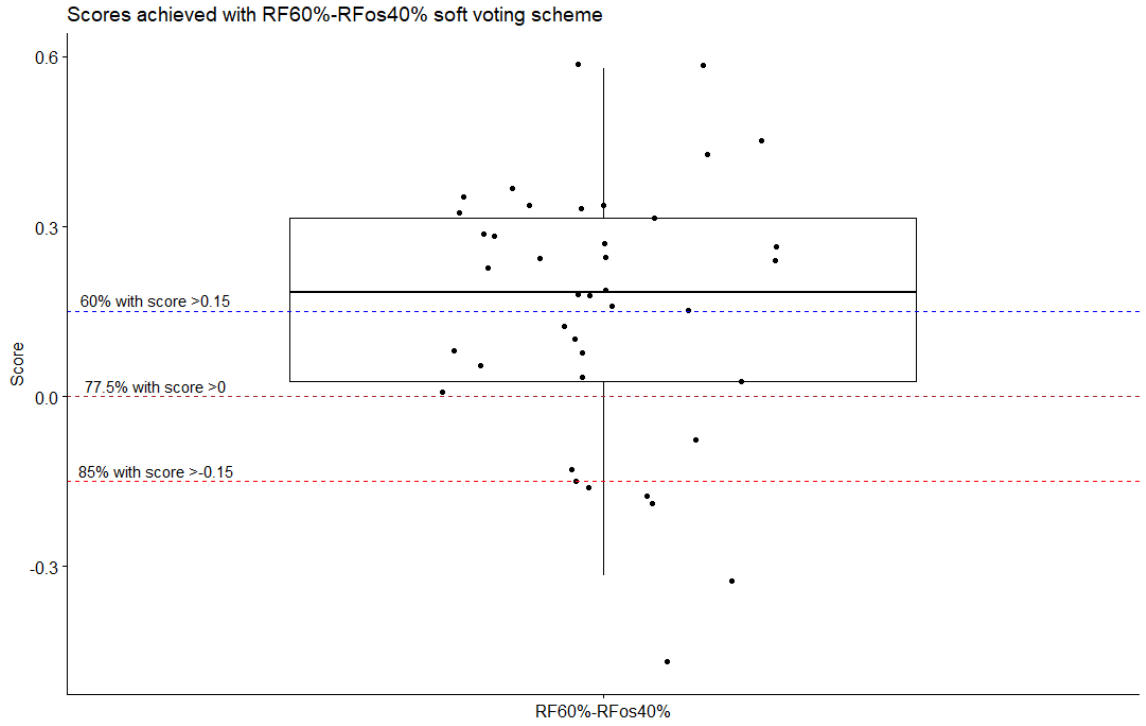


Figure 49: Classification scores achieved with the RF60%-RF_os40% ensemble in the whole manual segmentation cohort in the 40 different train:test splits with a 0.25 ratio left for testing

As we can see in Figure 49 in 77.5% of the different train:test splits of the dataset the RF60%-RF_os40% ensemble achieved positive classification scores. As a way of visualizing these scores in terms of survival curves, Figure 50 and Figure 51 demonstrate survival results for the patients stratified as high risk versus those stratified as low risk when the scoring result was nearly 0 (Figure 50) and when the scoring result was even slightly negative (Figure 51).

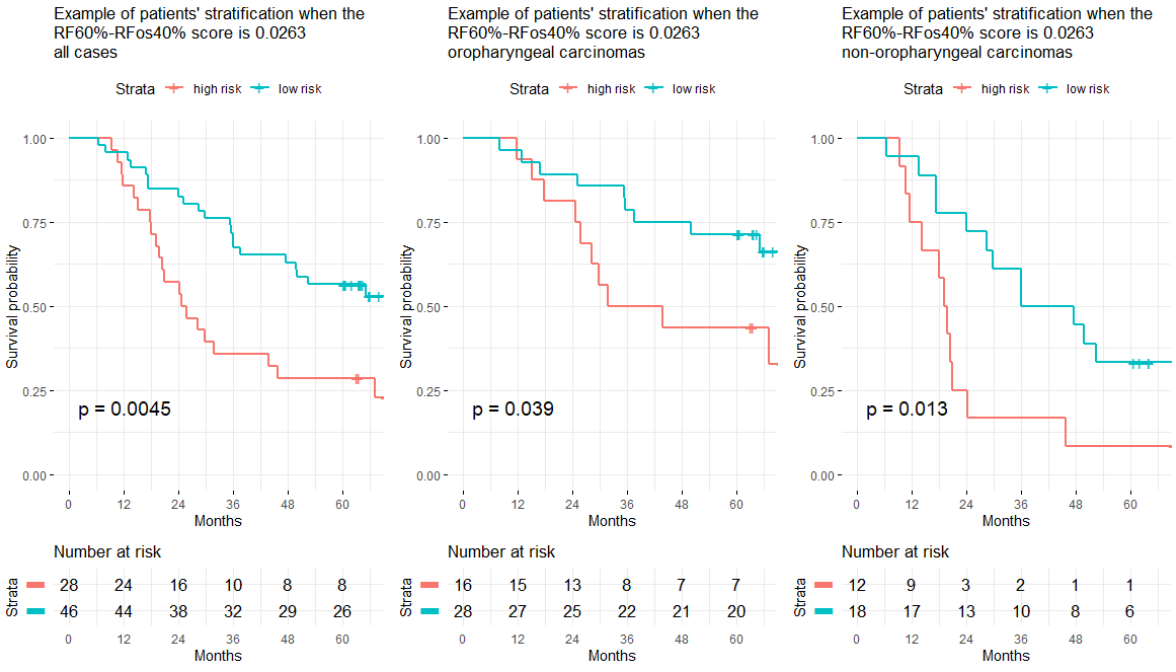


Figure 50: Example of patients' stratification with RF60%-RF_os40% ensemble's classification score equal to 0.0263 (all cases, OPC, non-OPC)

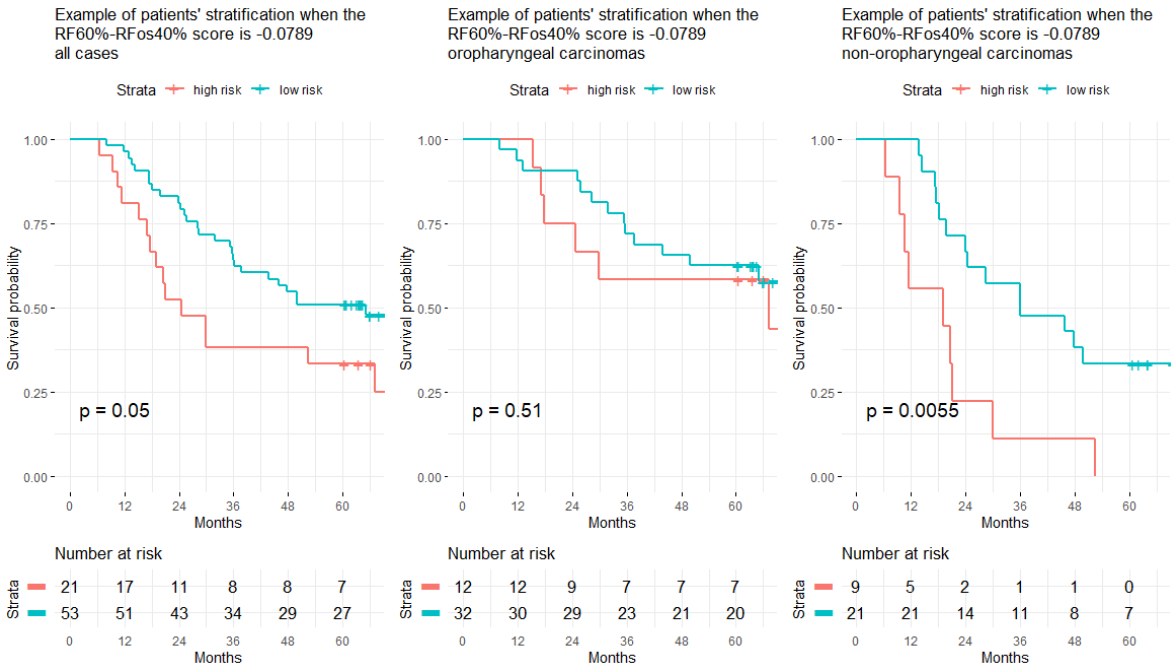


Figure 51: Example of patients' stratification with RF60%-RF_os40% ensemble's classification score equal to -0.0789 (all cases, OPC, non-OPC)

Stratification results in those two examples were statistically important ($p<0.05$) in both cases in the non-oropharyngeal subgroup but no difference in the survival curves was found in the oropharyngeal group in the case of the negative scoring result (Figure 50). Summing up, overall results were quite encouraging for the non-OPC subgroup of patients. Nonetheless the fact that there is a very large proportion of patients in our dataset with oropharyngeal carcinomas and with unknown HPV status emerged as a serious limitation. Therefore, we decided to proceed training in the case of the auto-segmentation cohort by applying exclusion criteria in the group of patients with oropharyngeal carcinomas.

3.2 Results on auto-segmented CT images

3.2.1 Unsupervised learning results on auto-segmented CT images

As we will observe in the following Figures: 52-59, unsupervised learning with GMMs achieved better classification results when all features were taken into consideration and OPC patients without premature death were excluded. Moreover, we observed that OPC patients without premature death tend to be widespread in the two-dimensional space, after dimensionality reduction (Figure 56) . Therefore, we believe that inclusion of such widespread cases will complicate the learning process when training the different classifiers. While interpreting these Figures we should also beware of the fact that accuracy is over-estimated when patients with oropharyngeal squamous cell carcinoma (OPSCC) without premature death are not excluded, because of the resulting very large proportion of patients without premature death in those cohorts. Regarding the results when using only "robust" features significantly associated ($p\text{-value}<0.05$) or with a trend of association ($p\text{-value}<0.1$) with survival, we found these results to be inferior (Figures

58, 59) to the results when all features were included (Figure 57) before the dimensionality reduction. Finally, as we can observe in Figures 52 and 53 when excluding patients with OPSCC without premature death, there seem to be a 3-cluster tendency in our dataset (one cluster with overrepresentation of low-risk patients and two clusters with overrepresentation of high-risk patients).

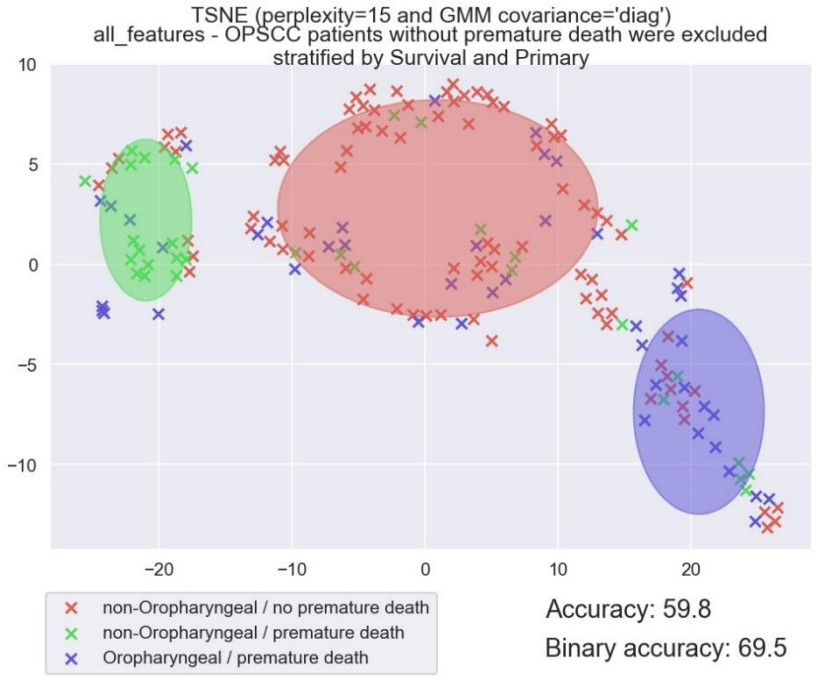


Figure 52: Unsupervised clustering (GMM covariance= "diag", 3 clusters) results in the cohort where patients with OPSCC without premature death were excluded ,and all features were kept before dimensionality reduction with t-SNE (perplexity=15)

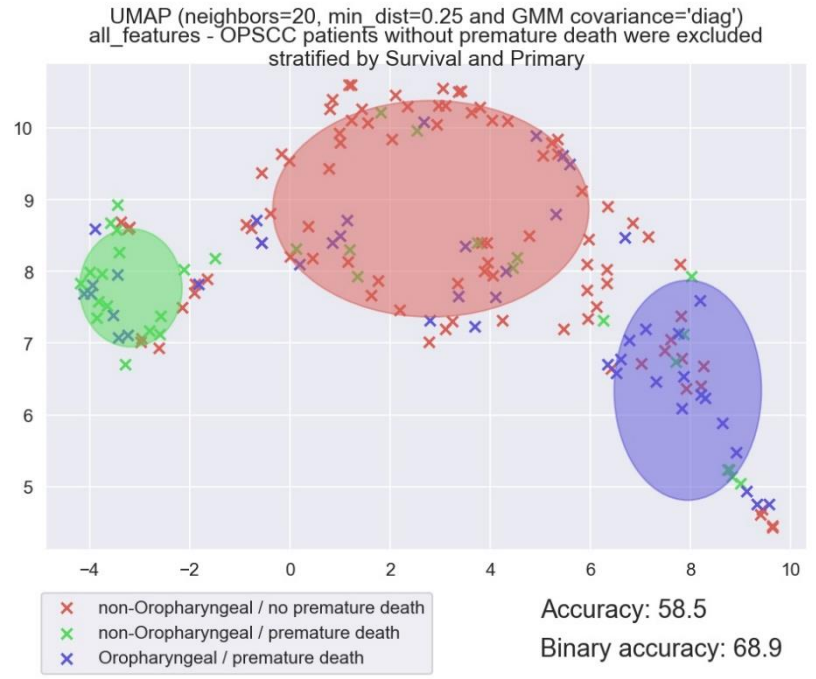


Figure 53: Unsupervised clustering (GMM covariance= "diag", 3 clusters) results in the cohort where patients with OPSCC without premature death were excluded ,and all features were kept before dimensionality reduction with UMAP (n_neighbors=20, min_dist=0.25)

Figure 54: Unsupervised clustering (GMM covariance= "tied", 2 clusters) results in the cohort where patients with OPSCC without premature death were excluded ,and all features were kept before dimensionality reduction with t-SNE (perplexity=20)

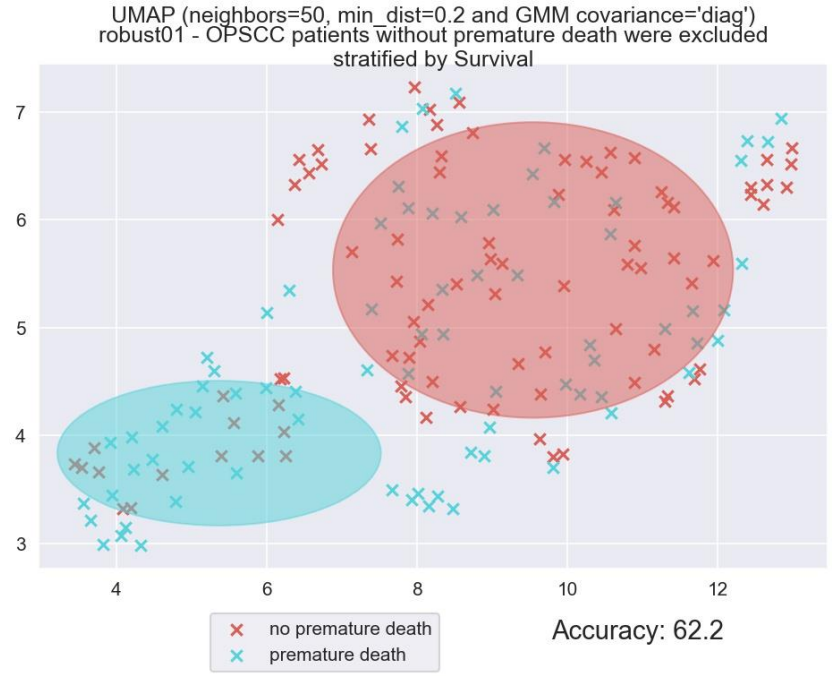


Figure 55: : Unsupervised clustering (GMM covariance= "diag", 2 clusters) results in the cohort where patients with OPSCC without premature death were excluded ,and only "robust" features with a trend of association (p-value<0.1) with survival were kept before dimensionality reduction with UMAP (n_neighbors=50, min_dist=0.2)

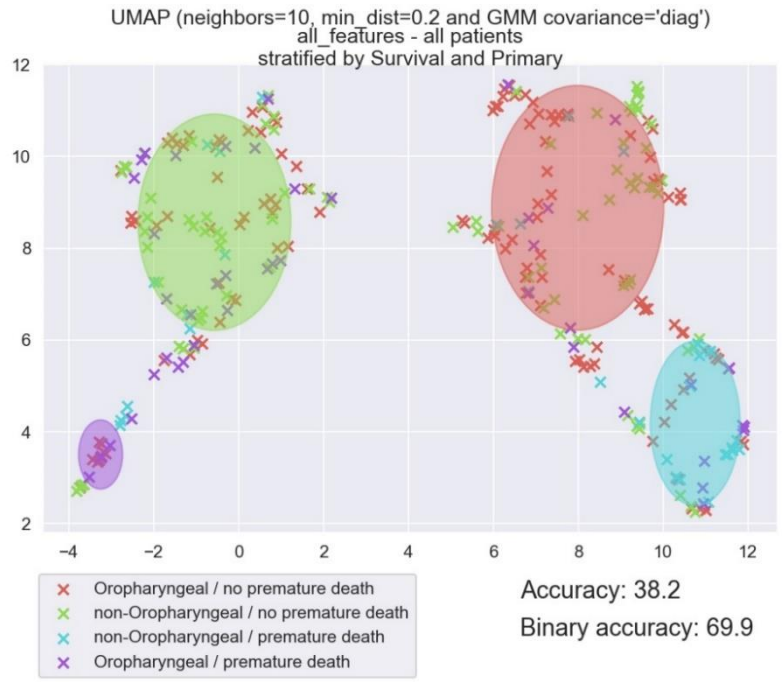


Figure 56: Unsupervised clustering (GMM covariance= "diag", 4 clusters) results in the cohort where all patients were included and all features were kept before dimensionality reduction with UMAP (n_neighbors=10, min_dist=0.2)

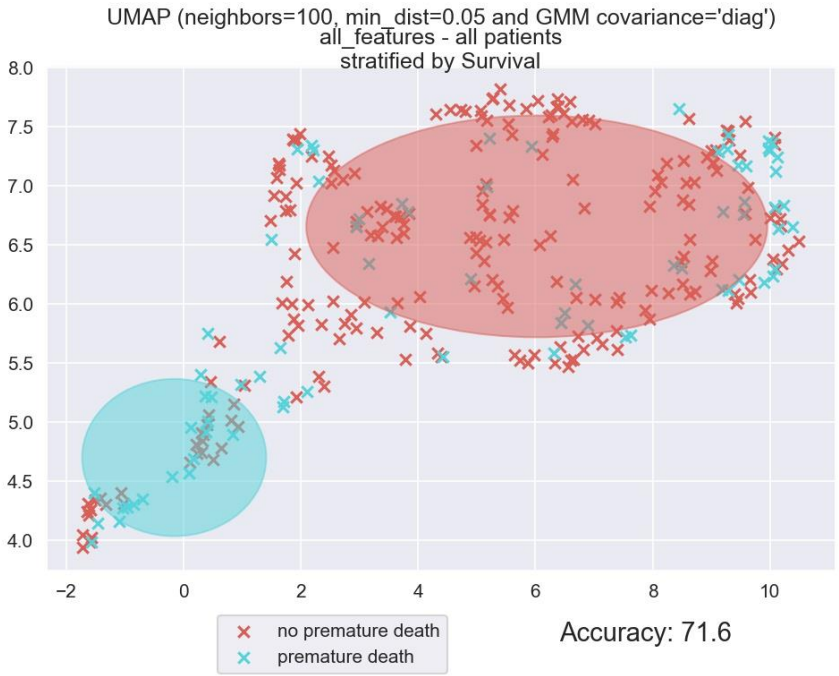


Figure 57: Unsupervised clustering (GMM covariance= "diag", 2 clusters) results in the cohort where all patients were included and all features were kept before dimensionality reduction with UMAP (n_neighbors=10, min_dist=0.2)

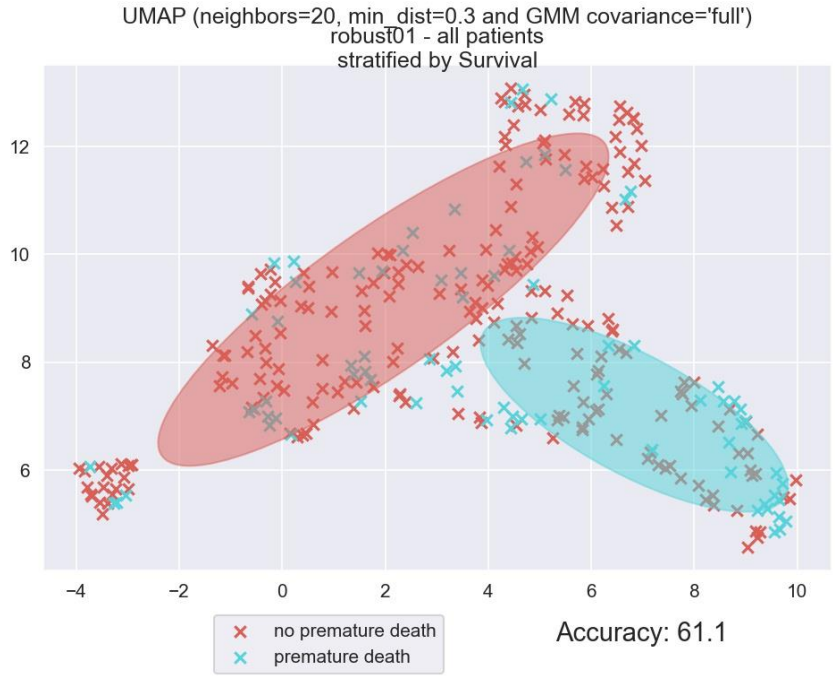


Figure 58: Unsupervised clustering (GMM covariance= "full", 2 clusters) results in the cohort where all patients were included ,and only "robust" features with a trend of association (p-value<0.1) with survival were kept before dimensionality reduction with UMAP (n_neighbors=20, min_dist=0.3)

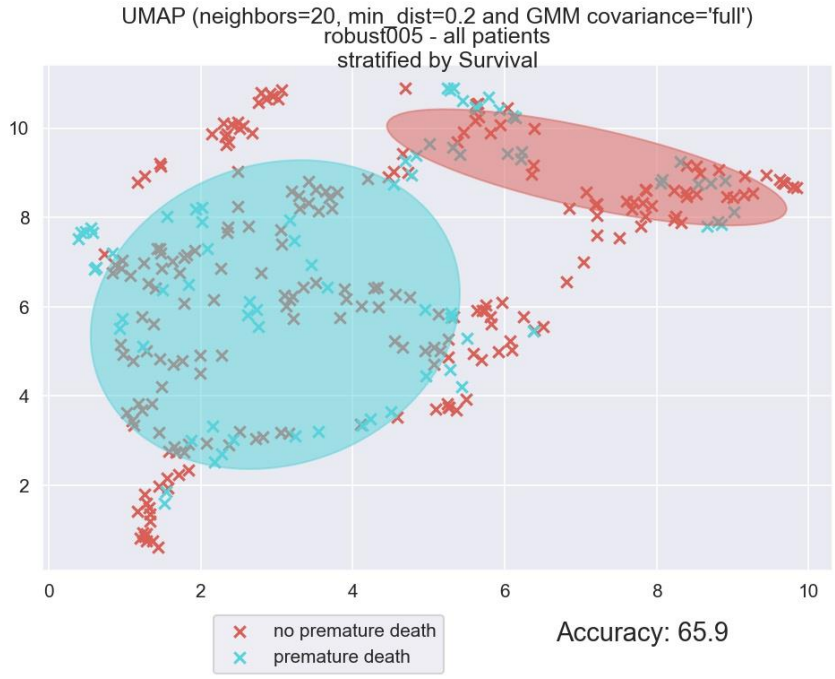


Figure 59: Unsupervised clustering (GMM covariance= "full", 2 clusters) results in the cohort where all patients were included ,and only "robust" features significantly associated (p-value<0.05) with survival were kept before dimensionality reduction with UMAP (n_neighbors=20, min_dist=0.2)

3.2.2 Supervised learning results on auto-segmented CT images

Supervised learning results will be presented separately for the two sub-cohorts created after the application of further exclusion criteria in the group of patients with oropharyngeal carcinomas (one with 47 patients where all OPC patients with OS > 5 years were excluded, and another one with 51 patients where 4 extra HPV-negative OPC patents with OS > 5 years were included; see Tables 12 and 13 in section "2.1 Material" for characteristics of the patients in each sub-cohort). PCA analysis will be presented prior to training results. We shall present results regarding two different ratios left for testing, 0.3 and 0.4. Dependence on the different dataset's train-test splits, of the variance explained shall also be investigated for the optimal number of components (Figures 61 and 85).

Classifiers were tested on the validation set consisted of the three images per patient that hadn't been used for training (Figure 41 – section 2.2.7). We averaged the three risks (either hard voting or soft voting derived from probabilities) from the different classifiers to obtain an average risk for premature death for each patient. Classifiers and ensemble models were then sorted out based on their performance in terms of minimum standard deviation of the optimal risk cut-off values derived from survival analysis, during the first 2.5 years, in the 40 different train-test splits of the dataset. Classifiers and ensemble models with the 8 lowest standard deviations of the optimal cut-off values have been evaluated with different metrics. For the evaluation of the various classifiers' final classification decision, we used the same cut-off point (both the mean and the median were tested), for all 40 different train-test splits of the dataset. In the following evaluation plots ensemble models will be named by the contributing classifiers and the name will be starting with either "pb" when soft voting with probabilities or "risk" when hard voting, all separated by underscore. For exmple pb_RF_QDA_GNB name will be used for soft voting ensemble model from equal contribution of RF's, GNB's and QDA's probabilities, and risk_lSVM_QDA for combined hard voting of the lSVM and the QDA.

3.2.2.1 Supervised learning results on auto-segmented CT images - 47 patients

The exclusion of all OPC patients with OS > 5 years resulted in a quite balanced (ratio of premature death: 0.4) sub-cohort of 47 patients (see Table 12).

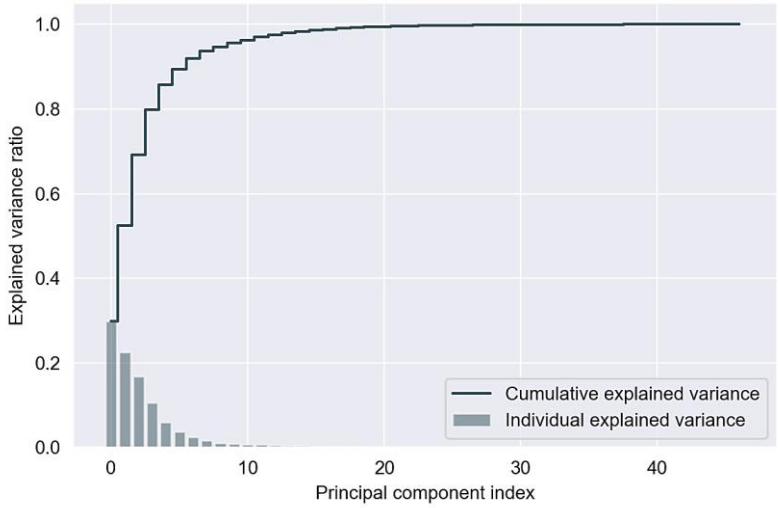


Figure 60: PCA analysis prior to training for the cohort with 47 patients

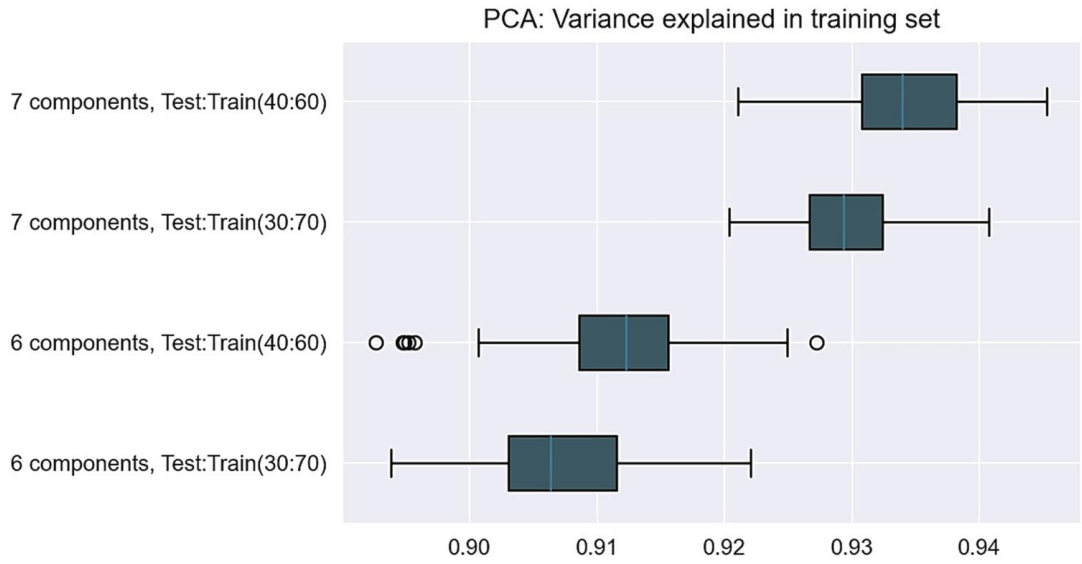


Figure 61: Variance explained in the training set by number of components and left out percentage for test, in the 40 different train-test splits of the dataset (cohort with 47 patients)

In the following subsections 3.2.2.1.a-d we will present training and validation results obtained when training with either 6 or 7 principal components and with a test ratio of either 0.3 or 0.4. The combination of 7 principal components and a test ratio of 0.3 seemed to be the most promising, from the variance analysis, and therefore the results obtained in that case, shall be presented in more details. In the other cases, we will focus on recall and f1 score, as we observed that classifiers with better performance in those two metrics achieved the best results regarding patients' overall survival related risk stratification.

3.2.2.1.a Training with 7 principal components and with a test ratio of 0.3

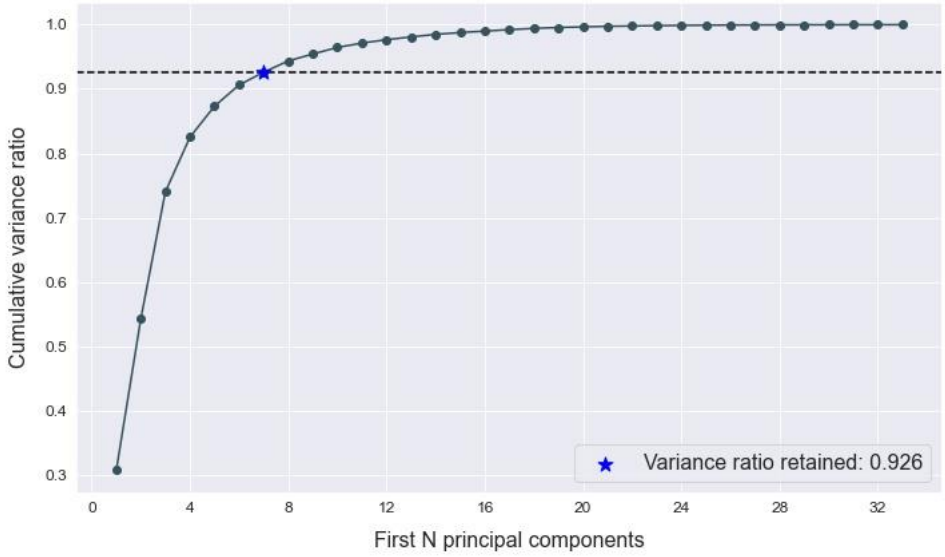


Figure 62: Example of variance explained in the training set in one of the 40 different train-test splits of the dataset

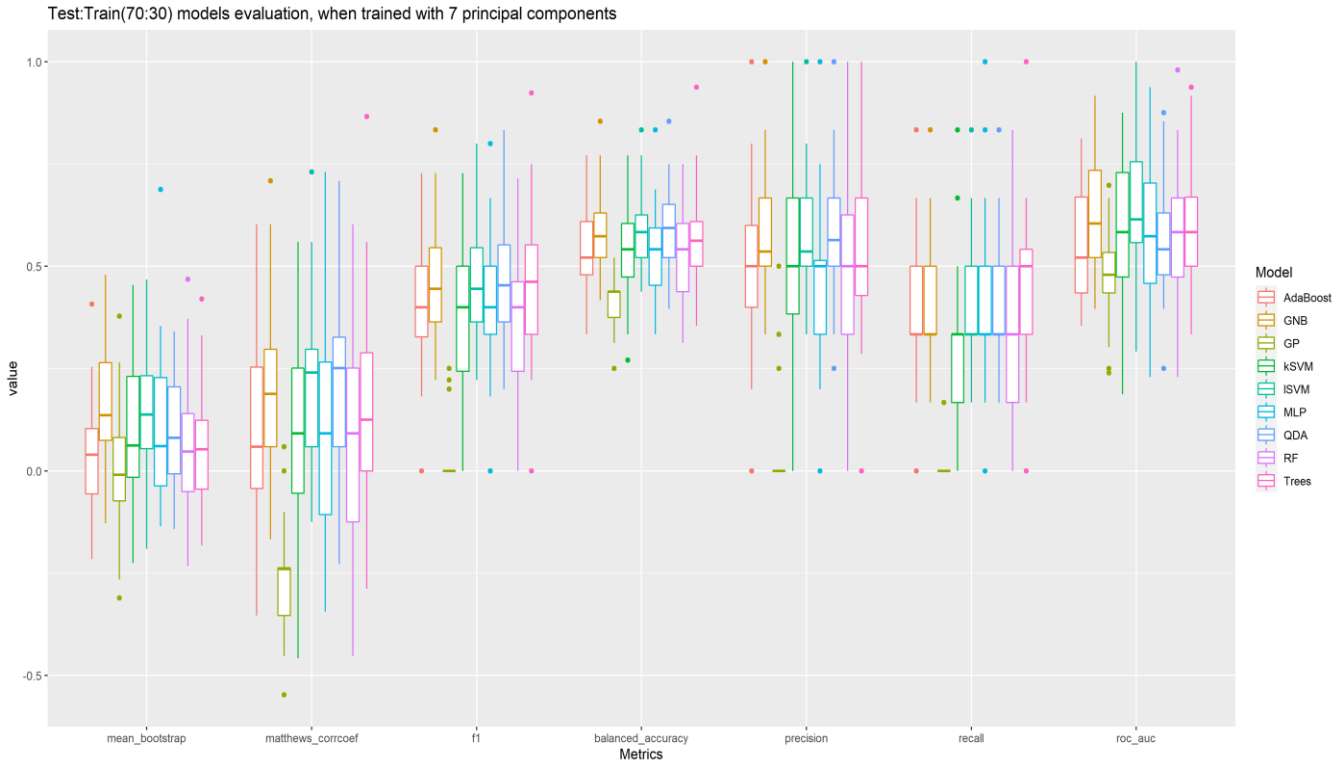


Figure 63: Training evaluation of different classifiers when trained with 7 principal components and with a test ratio of 0.3

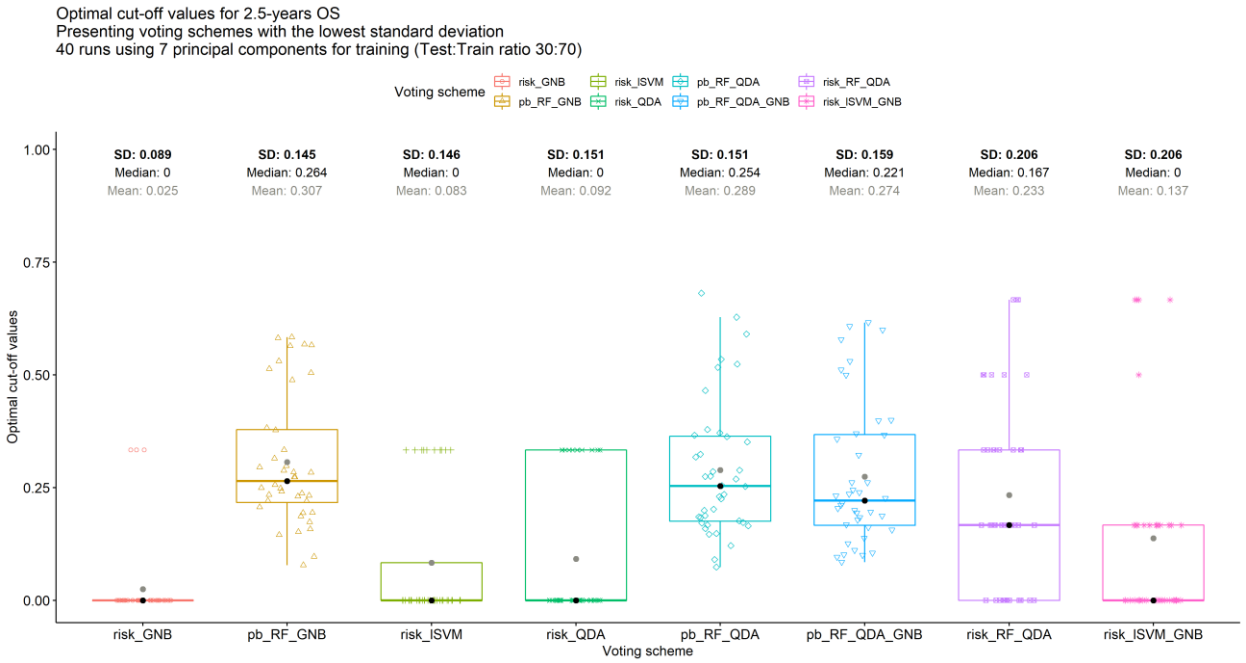


Figure 64: Optimal risk cut-off values in the 40 different train-test splits of the dataset, derived from survival analysis during the first 2.5 years. The 8 best performing, in terms of lowest standard deviation, classifiers and ensemble models (trained with 7 principal components and with a test ratio of 0.3) are being presented.

As we can observe in Figure 62 GNB, lSVM and QDA classifiers achieved the best results regarding MCC, balanced accuracy and precision, while GNB and lSVM remained the

better performing classifiers regarding bootstrapping and ROC AUC score. Moreover, GNB, lSVM and QDA remained within the 8 classifiers and ensemble models with the lowest standard deviation in terms of optimal cut-off values for patients' risk stratification (Figure 64). These three classifiers along with RF were also the base of the better performing ensemble models in the validation set (Figures 65-71). In Figures 65-71 various metrics' results achieved both when using the mean and when using the median of all 40 individual optimal cut-off values, will be compared.

Accuracy

Using mean as cut-off value for patients' risk stratification

Using median as cut-off value for patients' risk stratification

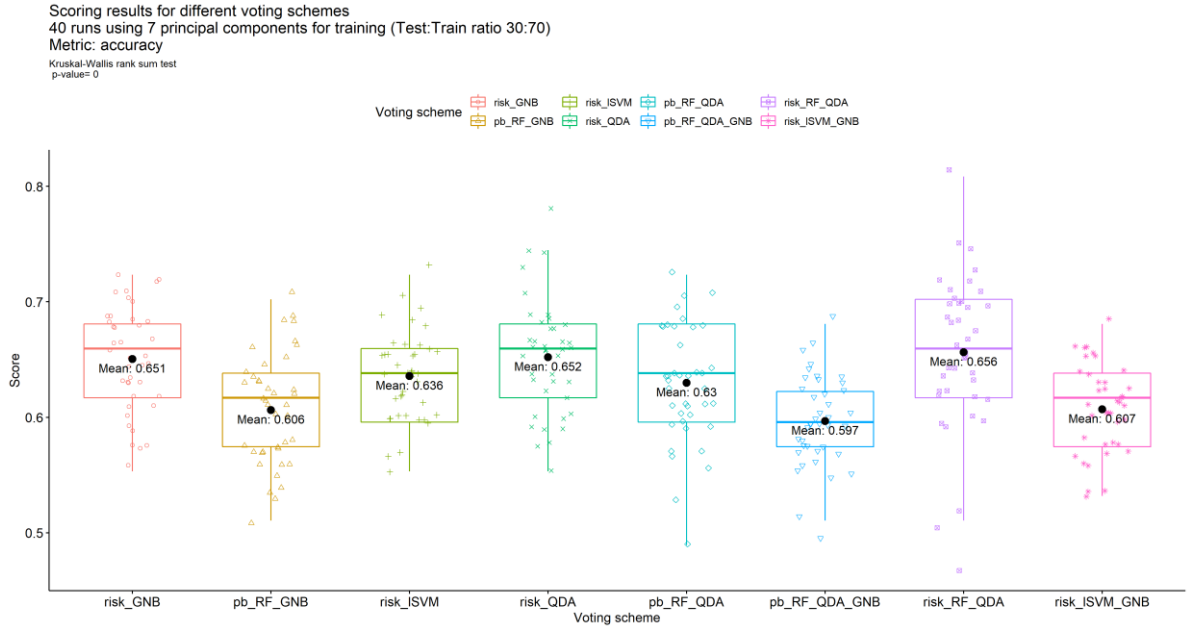
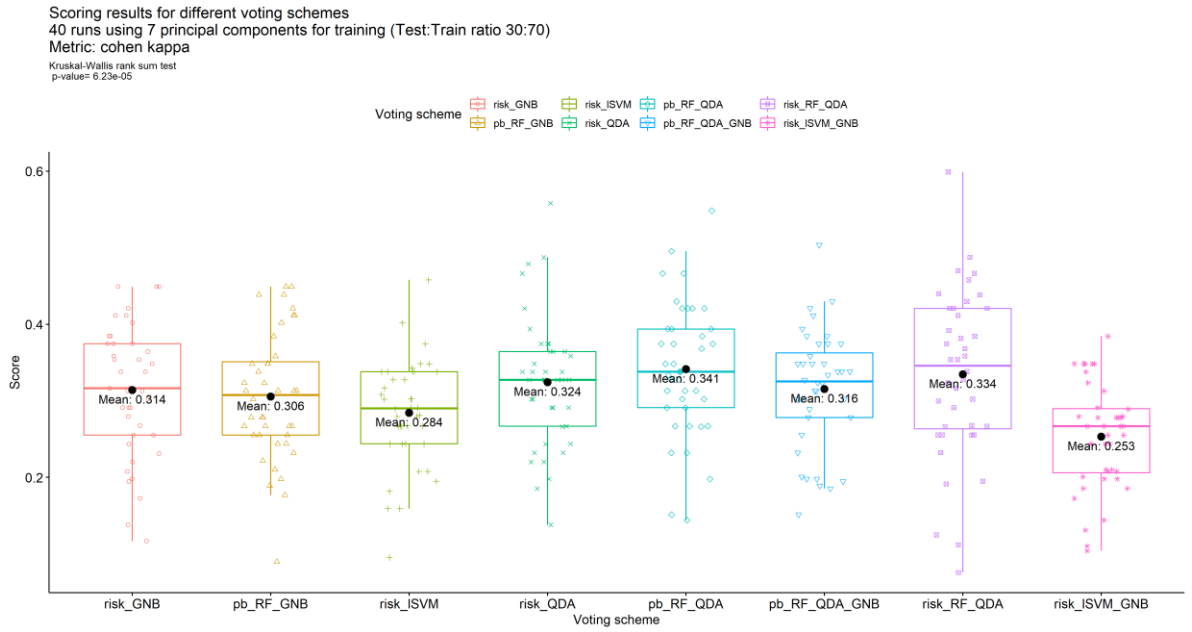


Figure 65: Accuracy results in the validation set for classifiers and ensemble models trained with 7 principal components and with a test ratio of 0.3

Regarding accuracy we observed that the usage of the mean led to better results for three classifiers (pb_RF_GNB, pb_RF_QDA and pb_RF_QDA_GNB), whereas the results of the other 5 classifiers remained the same (Figure 65).

Cohen's kappa

Using mean as cut-off value for patients' risk stratification



Using median as cut-off value for patients' risk stratification

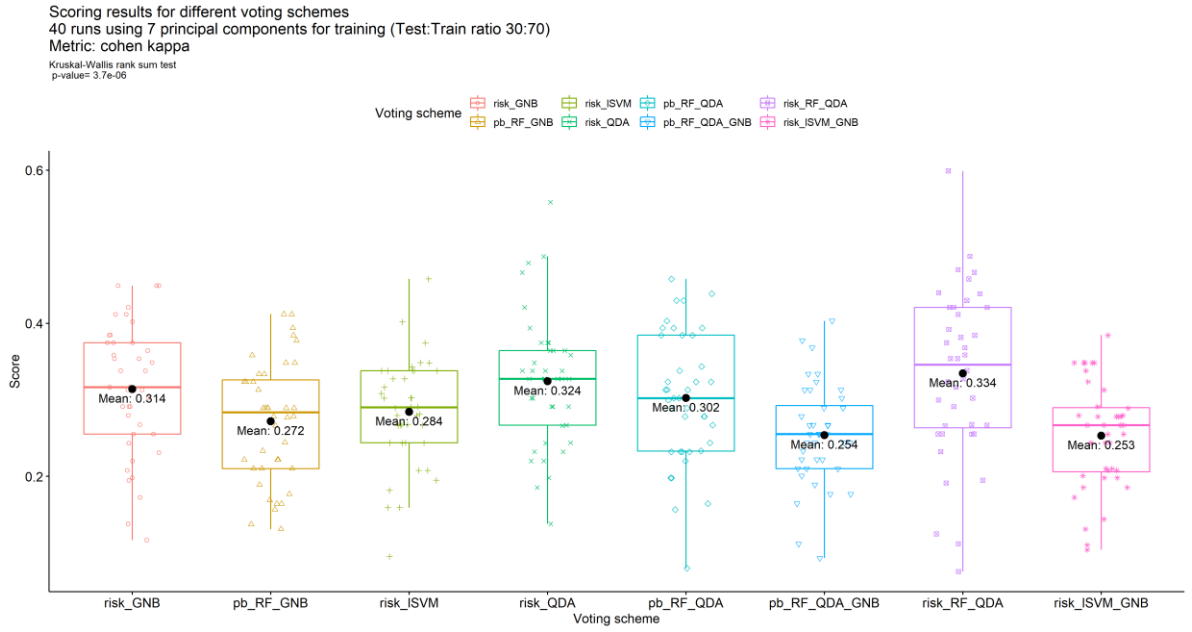
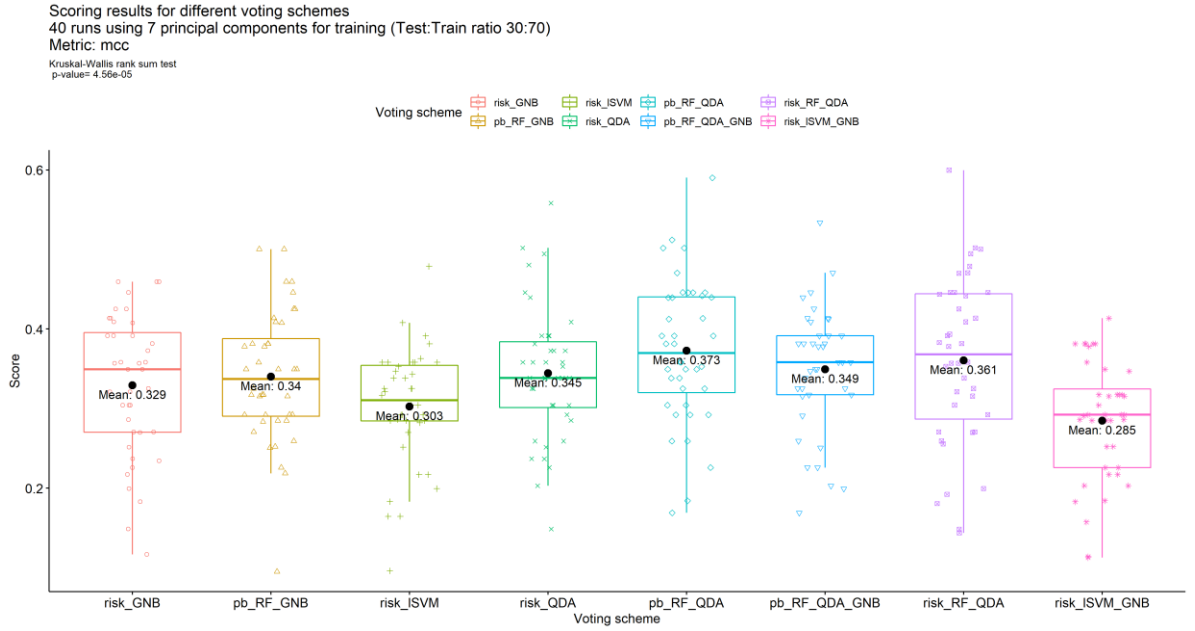


Figure 66: Cohen's kappa results in the validation set for classifiers and ensemble models trained with 7 principal components and with a test ratio of 0.3

Regarding both Cohen's kappa, MCC and precision, the usage of mean achieved again better results in the same three classifiers (pb_RF_GNB, pb_RF_QDA and pb_RF_QDA_GNB), while the rest had the same results (Figures 66, 67, 68).

MCC

Using mean as cut-off value for patients' risk stratification



Using median as cut-off value for patients' risk stratification

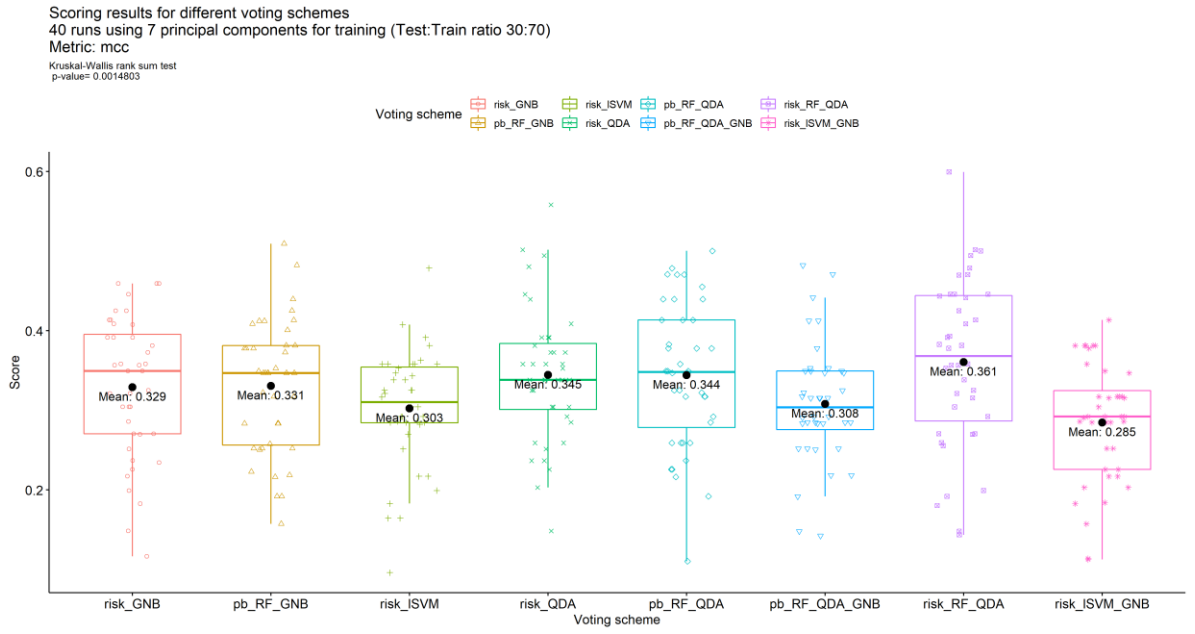
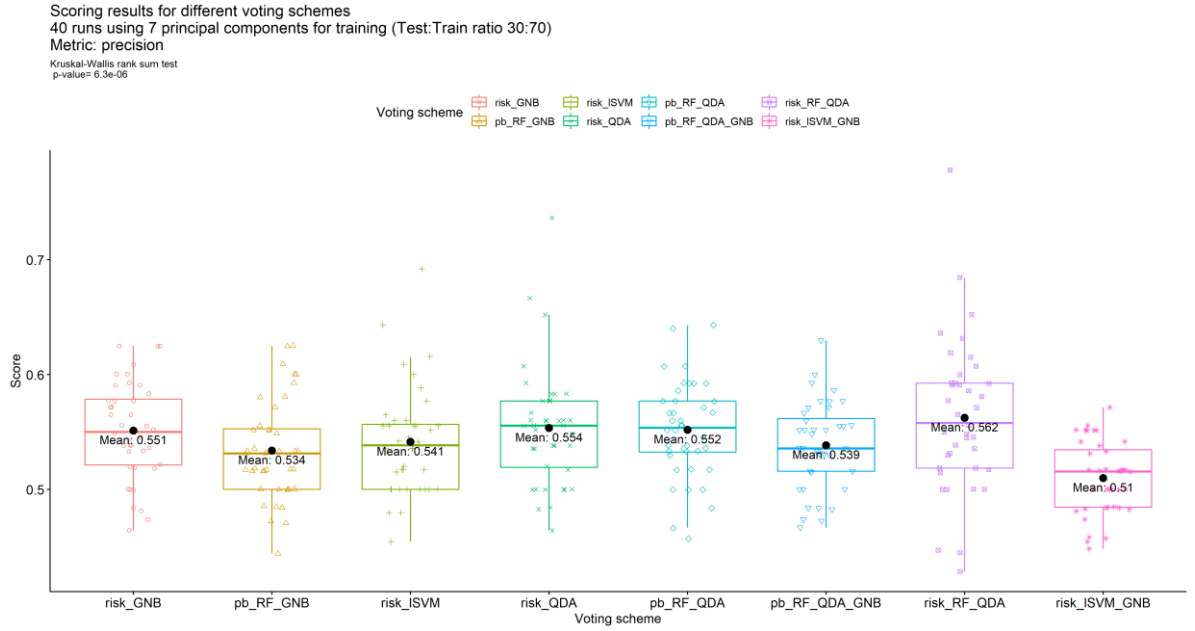


Figure 67: MCC results in the validation set for classifiers and ensemble models trained with 7 principal components and with a test ratio of 0.3

Precision

Using mean as cut-off value for patients' risk stratification



Using median as cut-off value for patients' risk stratification

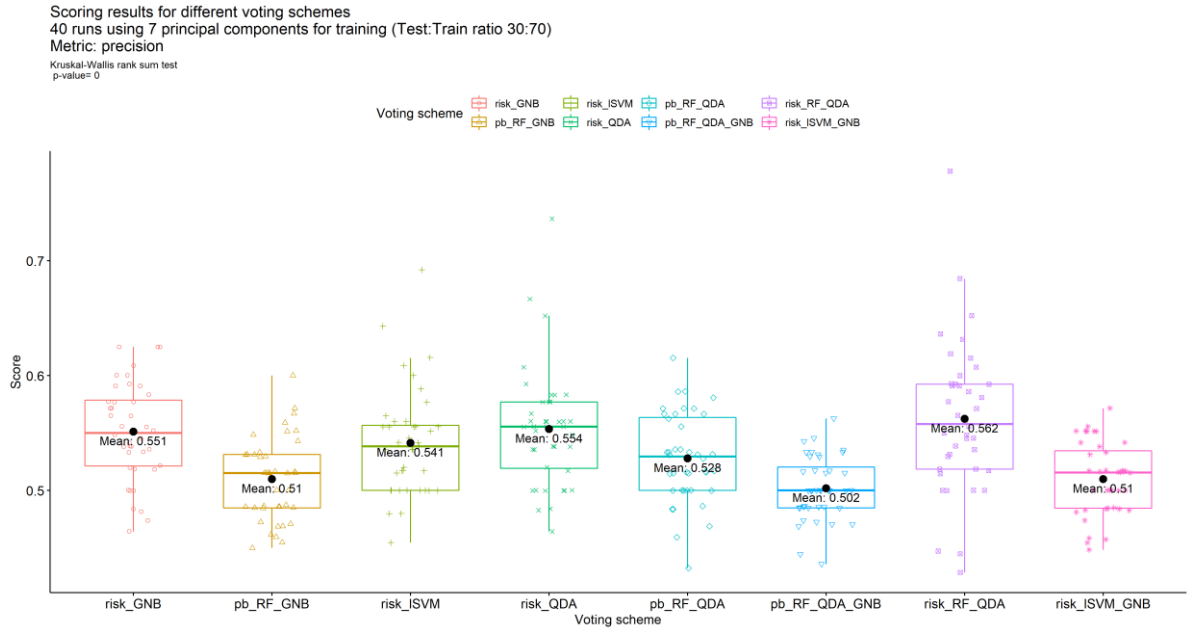
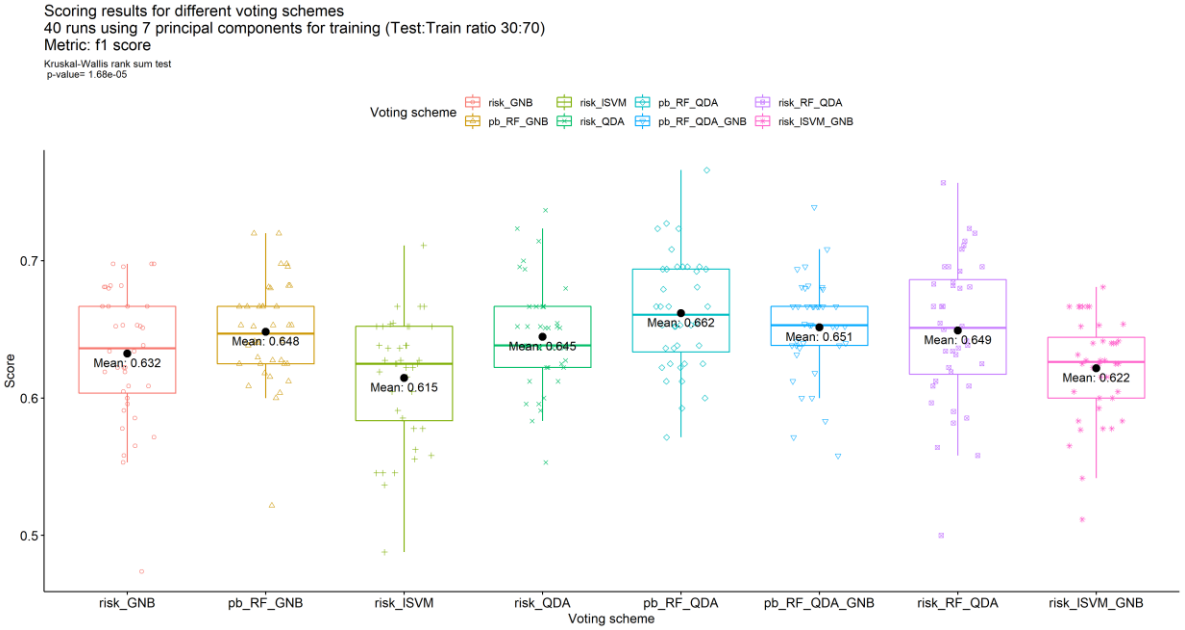


Figure 68: Precision results in the validation set for classifiers and ensemble models trained with 7 principal components and with a test ratio of 0.3

F1-score

Using mean as cut-off value for patients' risk stratification



Using median as cut-off value for patients' risk stratification

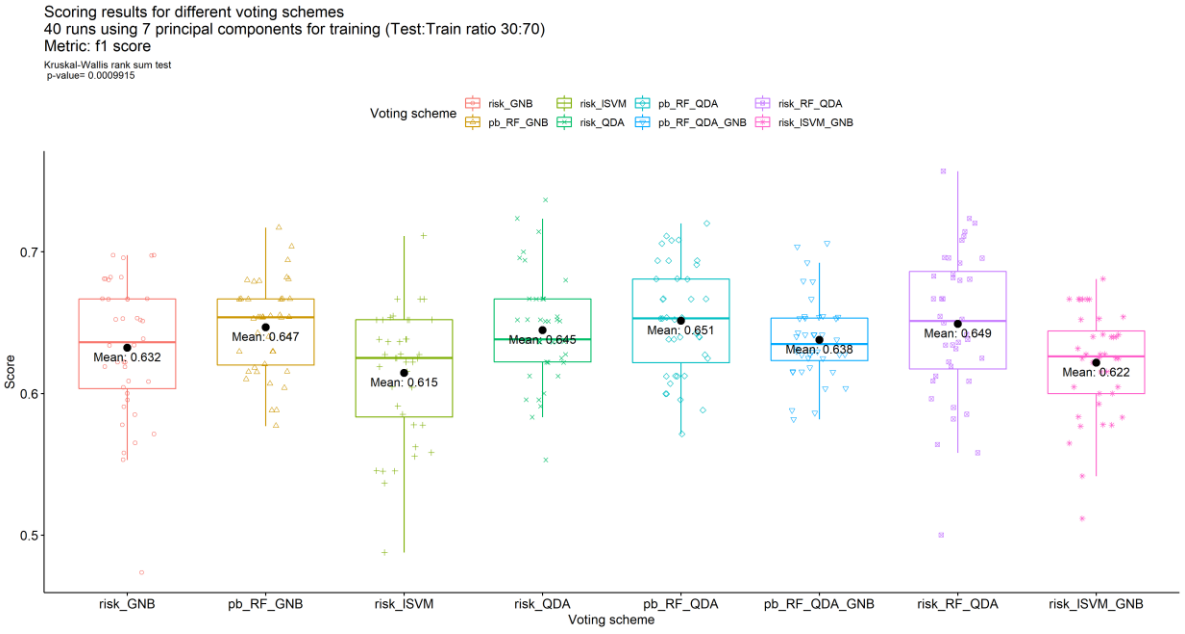
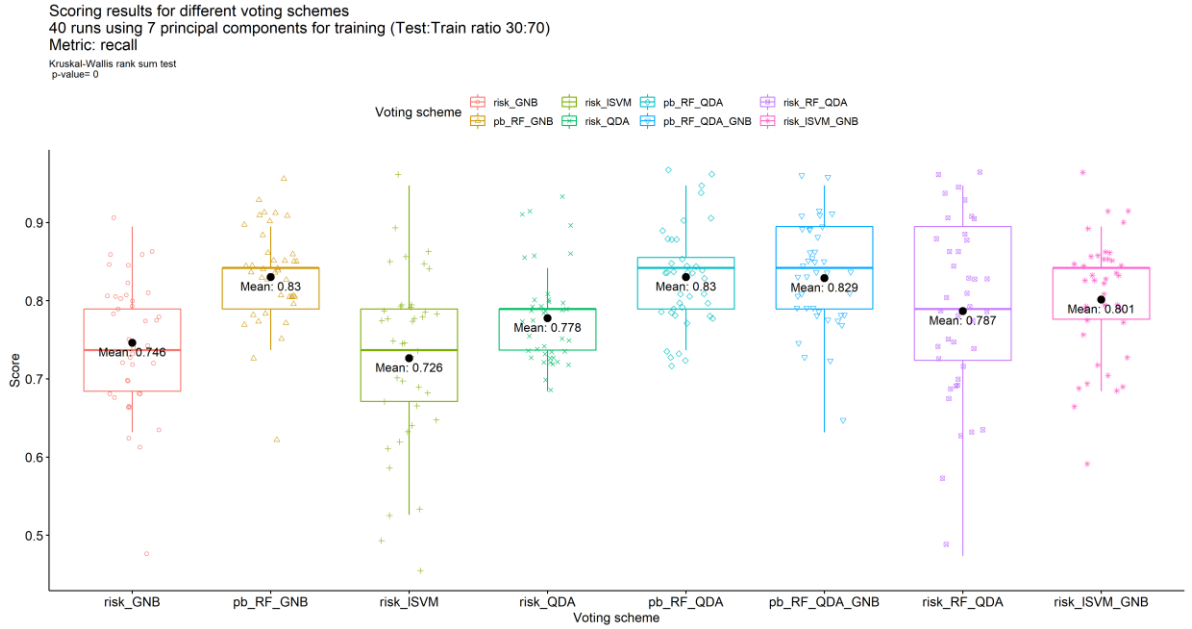


Figure 69: F1-score results in the validation set for classifiers and ensemble models trained with 7 principal components and with a test ratio of 0.3

Regarding F1-score the results were almost the same, and only slightly better in the same three classifiers (pb_RF_GNB, pb_RF_QDA and pb_RF_QDA_GNB) when using the mean (Figure 69).

Recall

Using mean as cut-off value for patients' risk stratification



Using median as cut-off value for patients' risk stratification

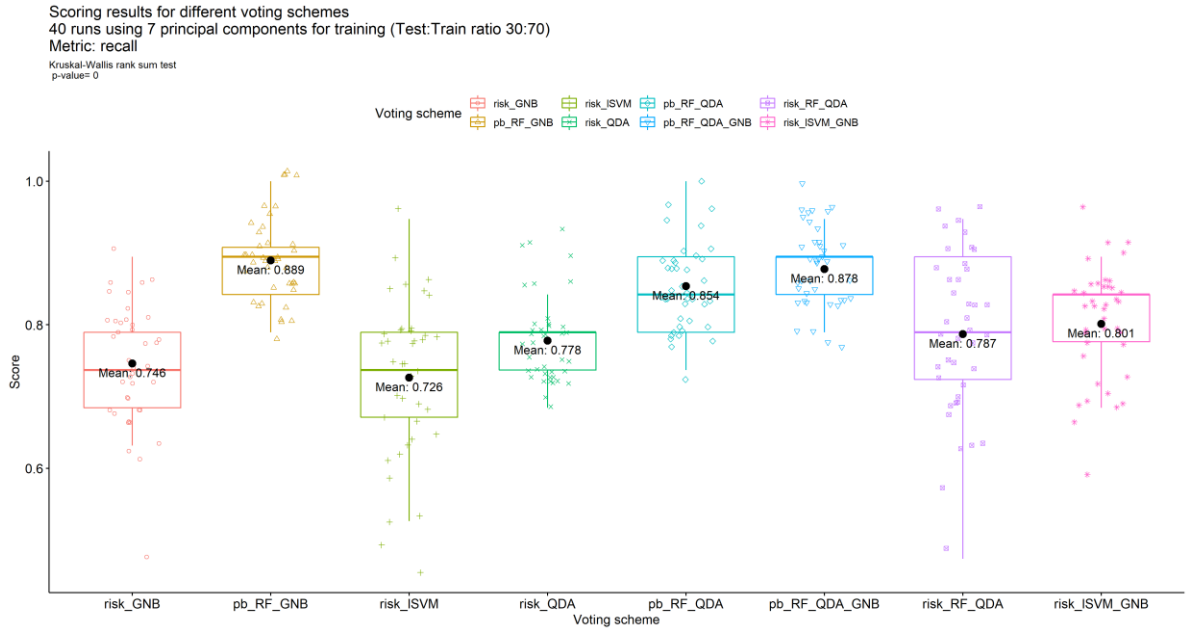


Figure 70: Recall results in the validation set for classifiers and ensemble models trained with 7 principal components and with a test ratio of 0.3

Regarding recall again the three ensemble models with probabilities soft voting (pb_RF_GNB, pb_RF_QDA and pb_RF_QDA_GNB) were differentiated between mean and median. However, in the case of recall the usage of median led to better results. Interestingly, in the case of recall risk_lSVM_GNB model emerged as the fourth best performing classifier.

The differences observed so far are in line with the fact that the usage of median led to lower cut-off values compared to mean, in the cases of soft voting ensemble models

(Figure 64). Lower cut-off values generally lead to more cases identified as high risk and therefore potentially increase false positives (resulting in lower precision) and decrease false negatives (resulting in higher recall), while F1-score's results that are equally affected by both recall and precision remain almost the same. Moreover, the custom metric whose scoring was favouring identifying high risk cases (scoring described in section "2.2.7 Evaluation metrics and validation"), showcased the same differences as recall (Figure 71).

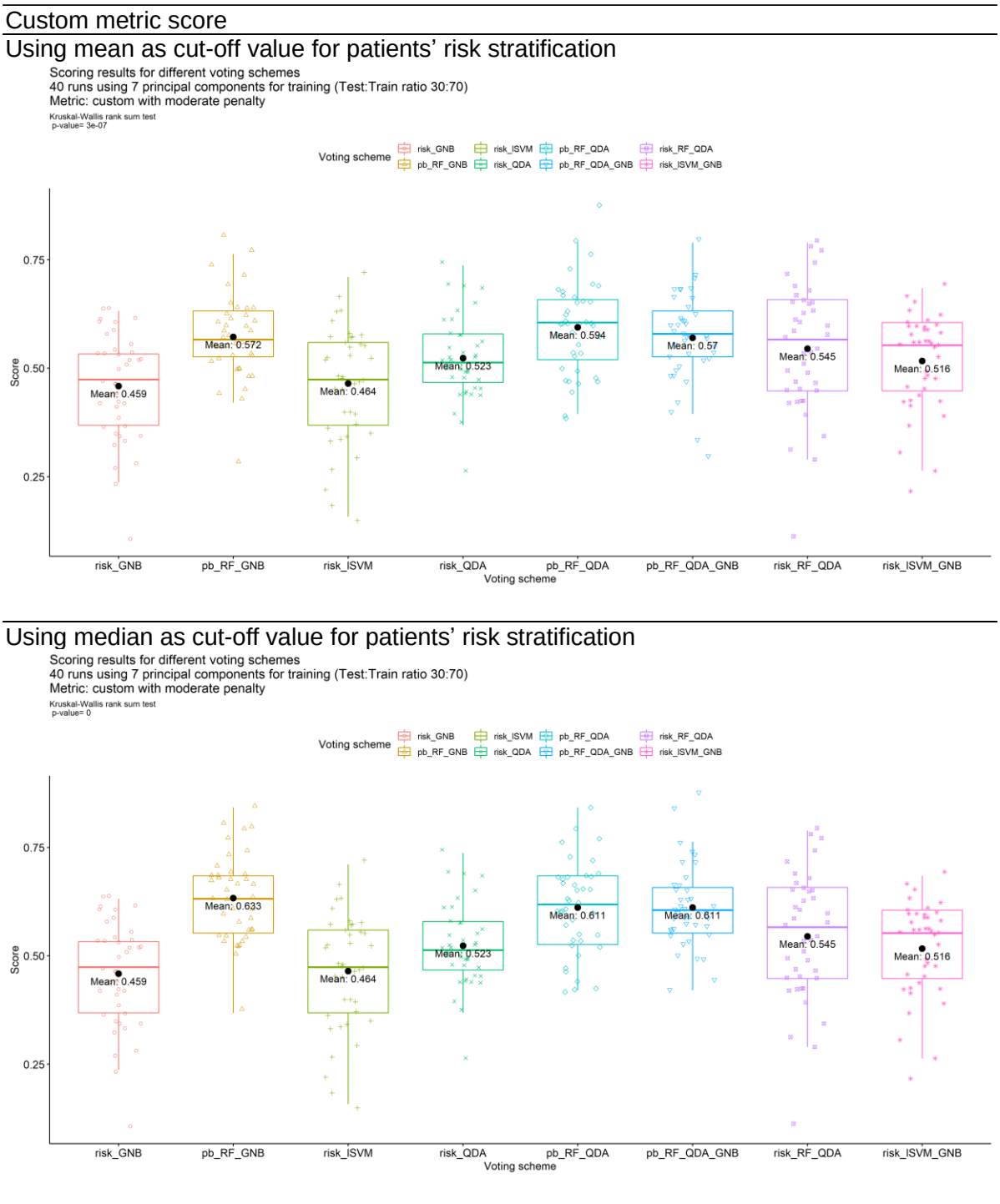


Figure 71: Custom metric results in the validation set for classifiers and ensemble models trained with 7 principal components and with a test ratio of 0.3

3.2.2.1.b Training with 6 principal components and with a test ratio of 0.3

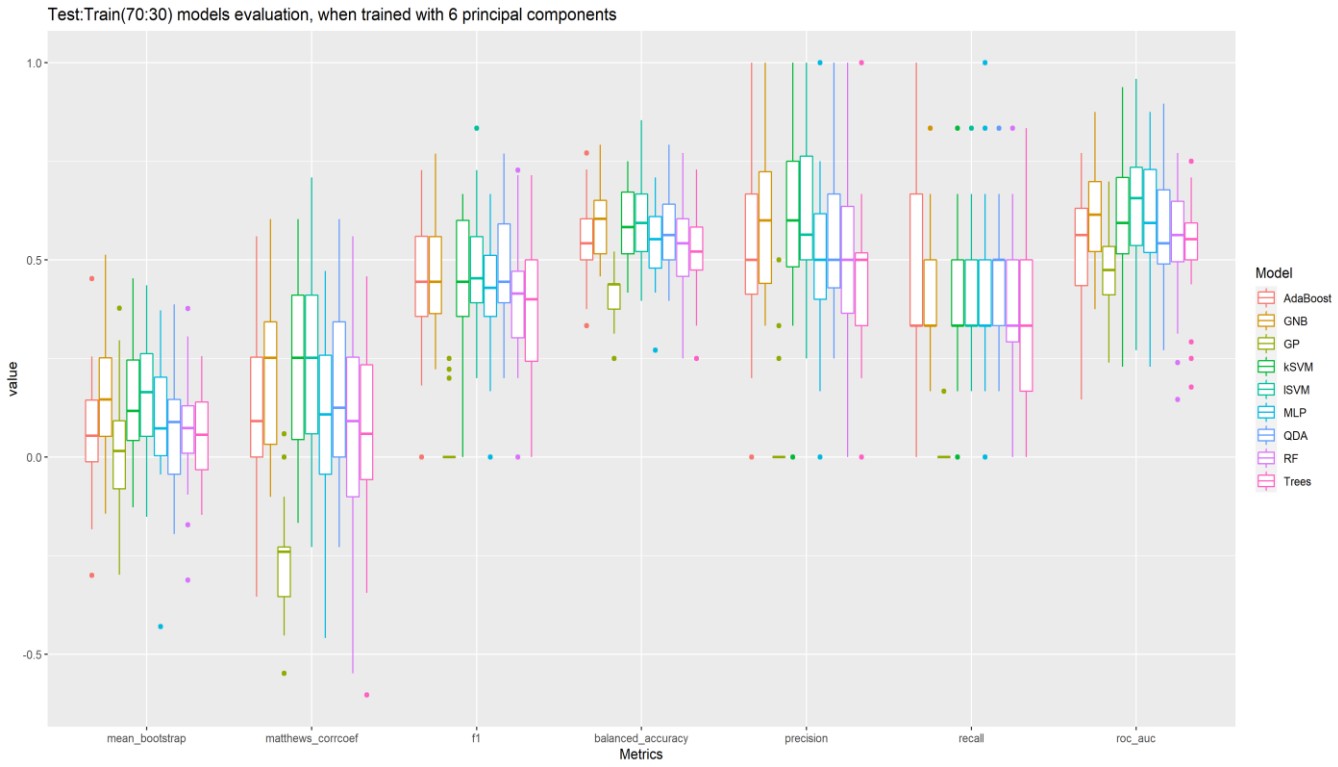


Figure 72: Training evaluation of different classifiers when trained with 6 principal components and with a test ratio of 0.3

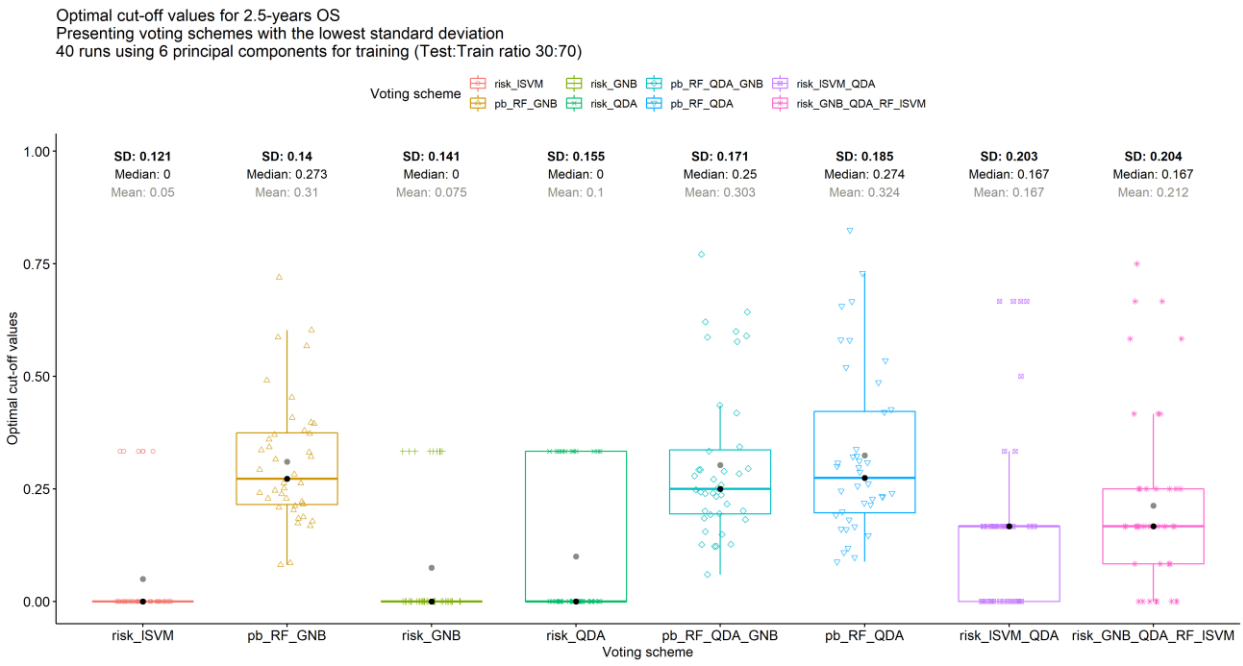
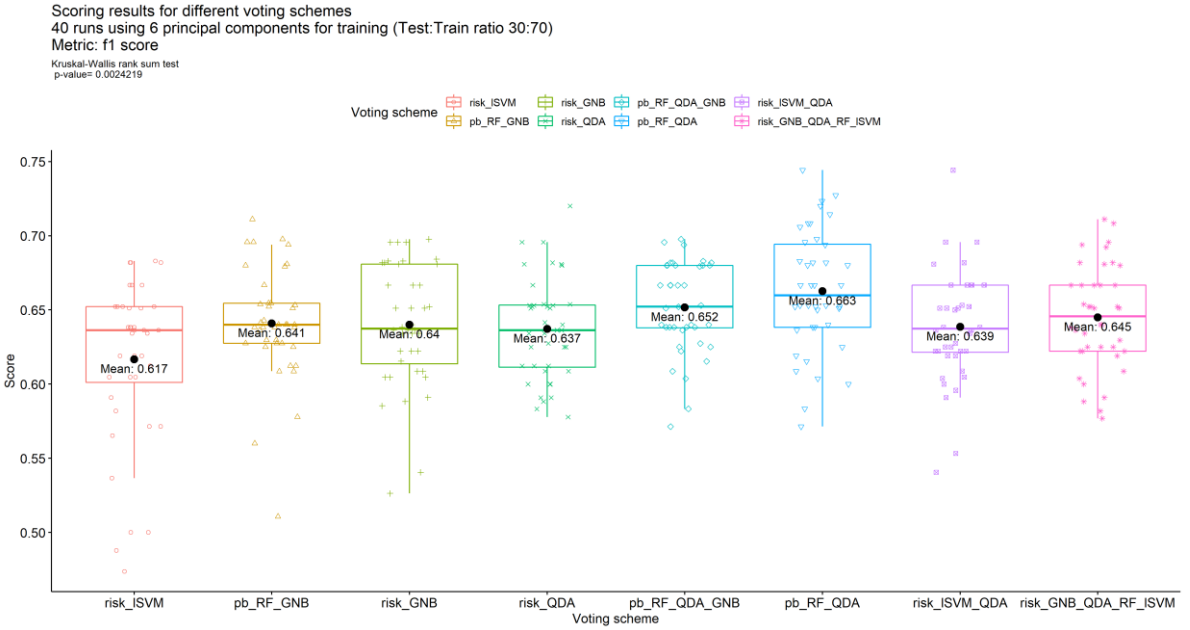


Figure 73: Optimal risk cut-off values in the 40 different train-test splits of the dataset, derived from survival analysis during the first 2.5 years. The 8 best performing, in terms of lowest standard deviation, classifiers and ensemble models (trained with 6 principal components and with a test ratio of 0.3) are being presented.

GNB and lSVM classifiers achieved the best results regarding MCC, balanced accuracy, ROC AUC score and bootstrapping when training with 6 principal components and with a test ratio of 0.3 (Figure 72), while GNB was the base of most of the better performing ensemble models in terms of lowest standard deviation of optimal risk cut-off values (Figure 73).

F1-score

Using mean as cut-off value for patients' risk stratification



Using median as cut-off value for patients' risk stratification

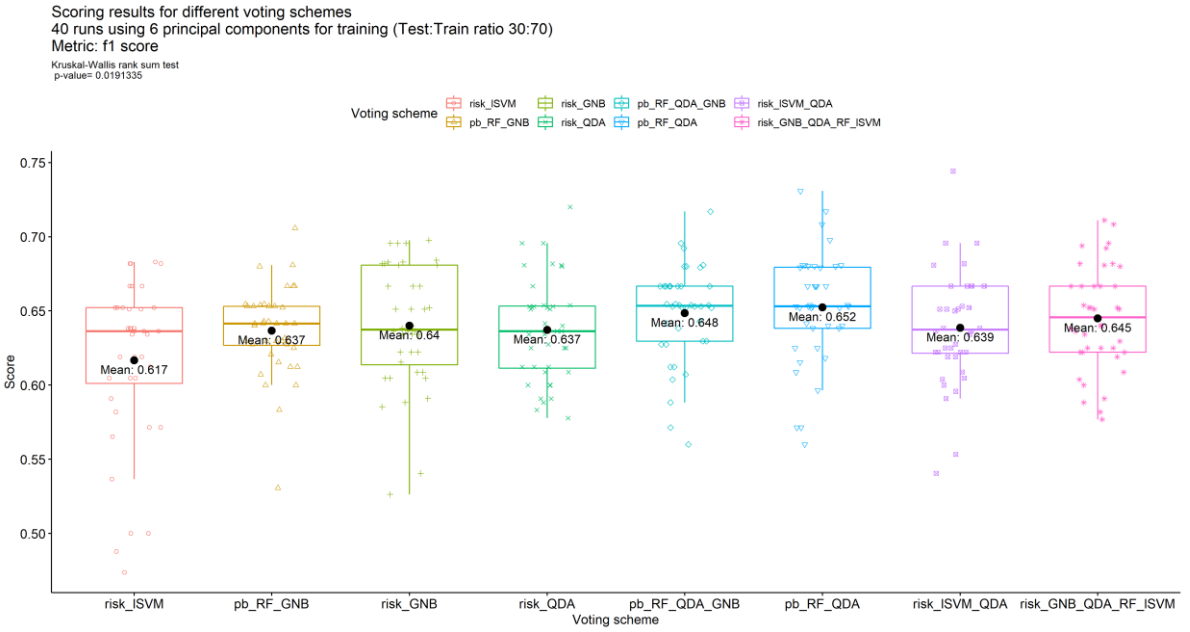
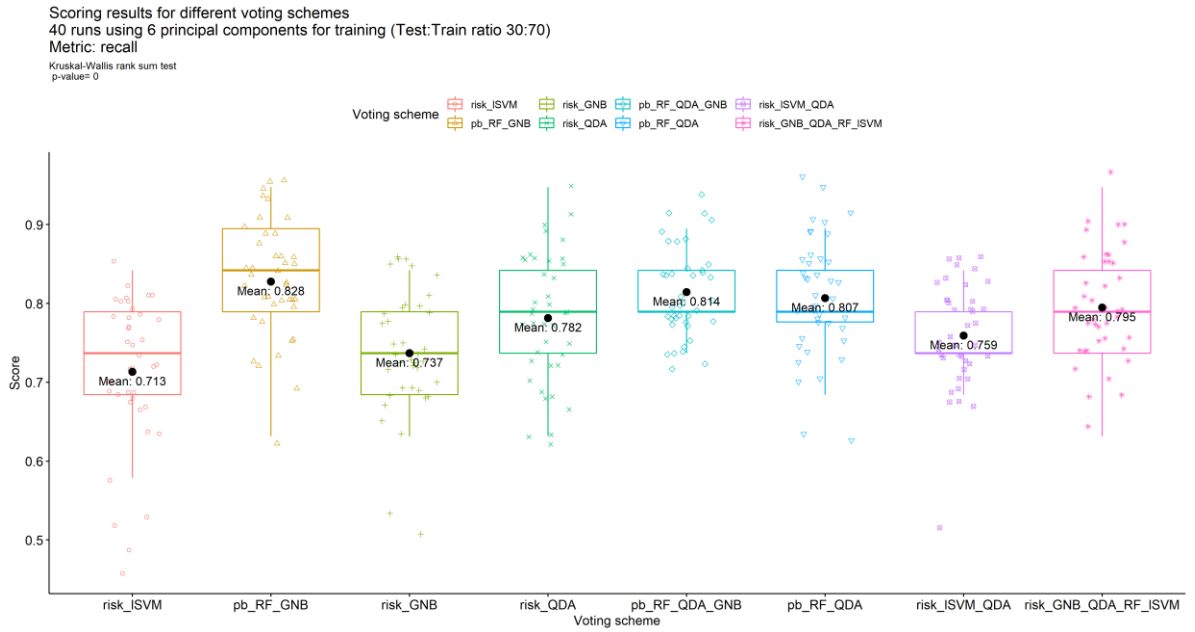


Figure 74: F1-score results in the validation set for classifiers and ensemble models trained with 6 principal components and with a test ratio of 0.3

Ensemble soft voting models pb_RF_GNB and pb_RF_QDA_GNB achieved the most promising results in the validation set, especially when using median as cut-off (Figures 73-75).

Recall

Using mean as cut-off value for patients' risk stratification



Using median as cut-off value for patients' risk stratification

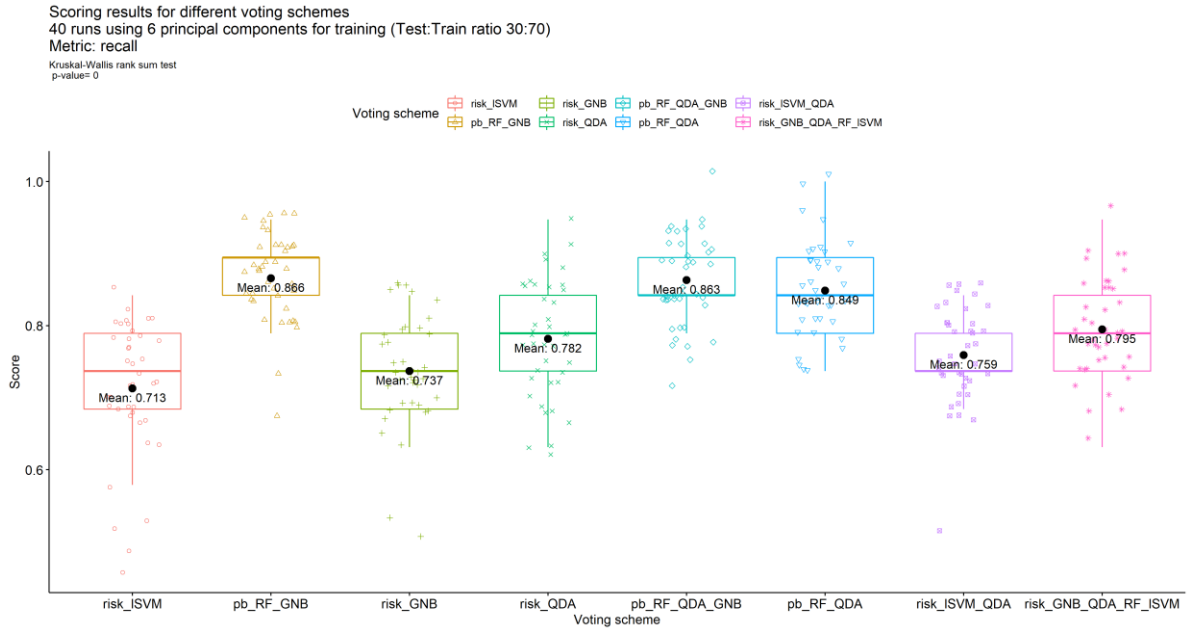


Figure 75: Recall results in the validation set for classifiers and ensemble models trained with 6 principal components and with a test ratio of 0.3

3.2.2.1.c Training with 6 principal components and with a test ratio of 0.4

In consistent with previous results, GNB and lSVM classifiers achieved again the best results regarding MCC, balanced accuracy, ROC AUC score and bootstrapping (Figure

76), while pb_RF_GNB and pb_RF_QDA_GNB were the better performing ensemble models in the validation set (Figures 77, 78 and 79).

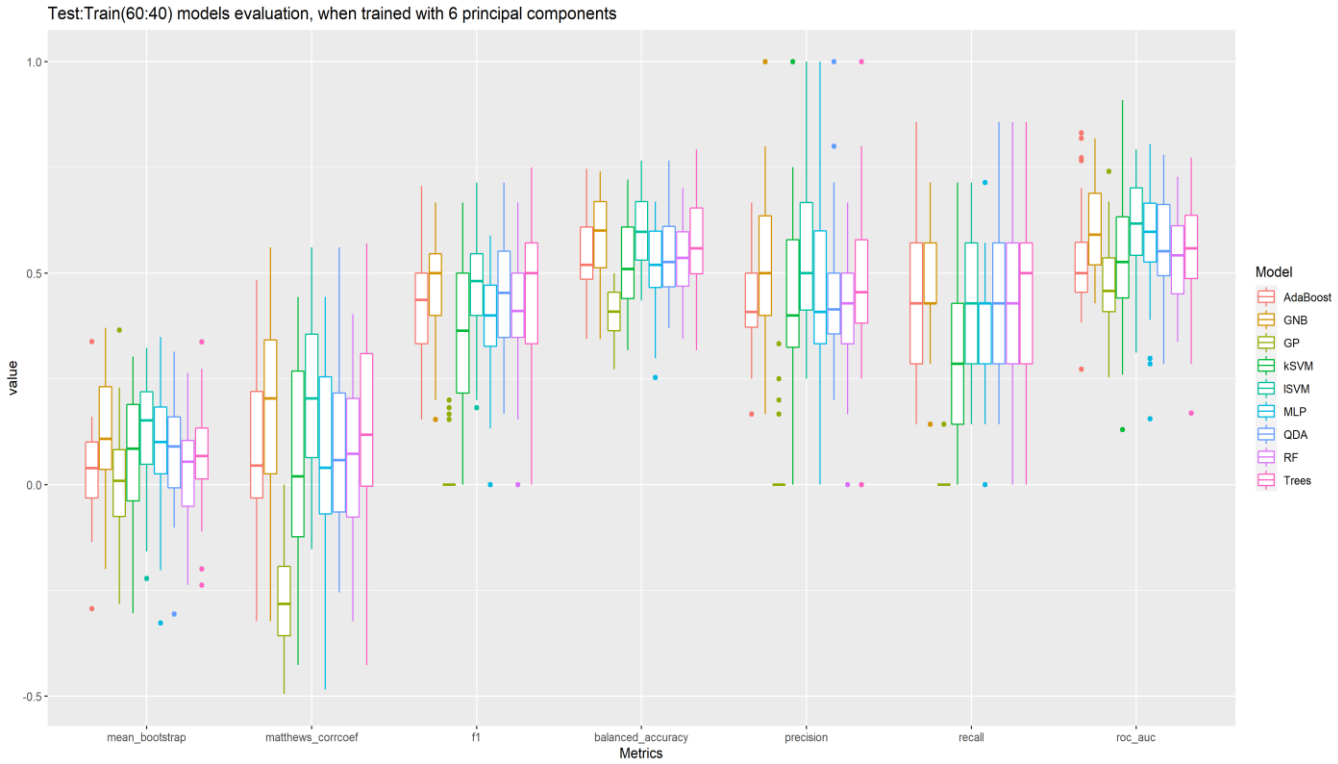


Figure 76: Training evaluation of different classifiers when trained with 6 principal components and with a test ratio of 0.4

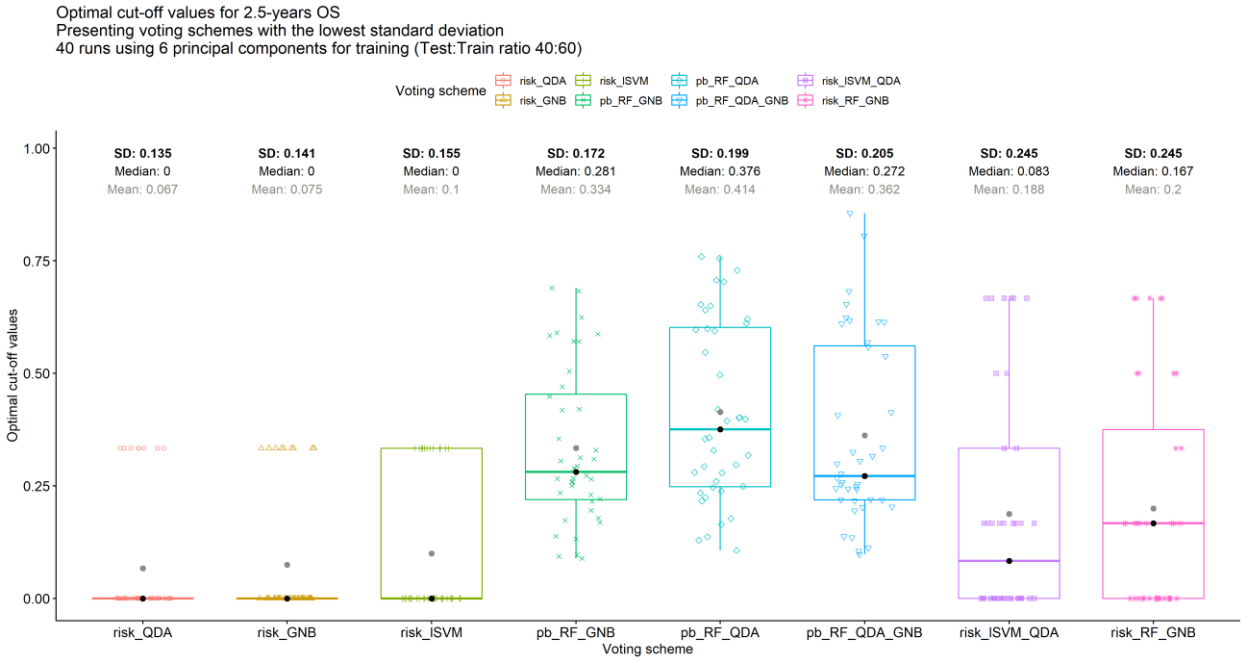
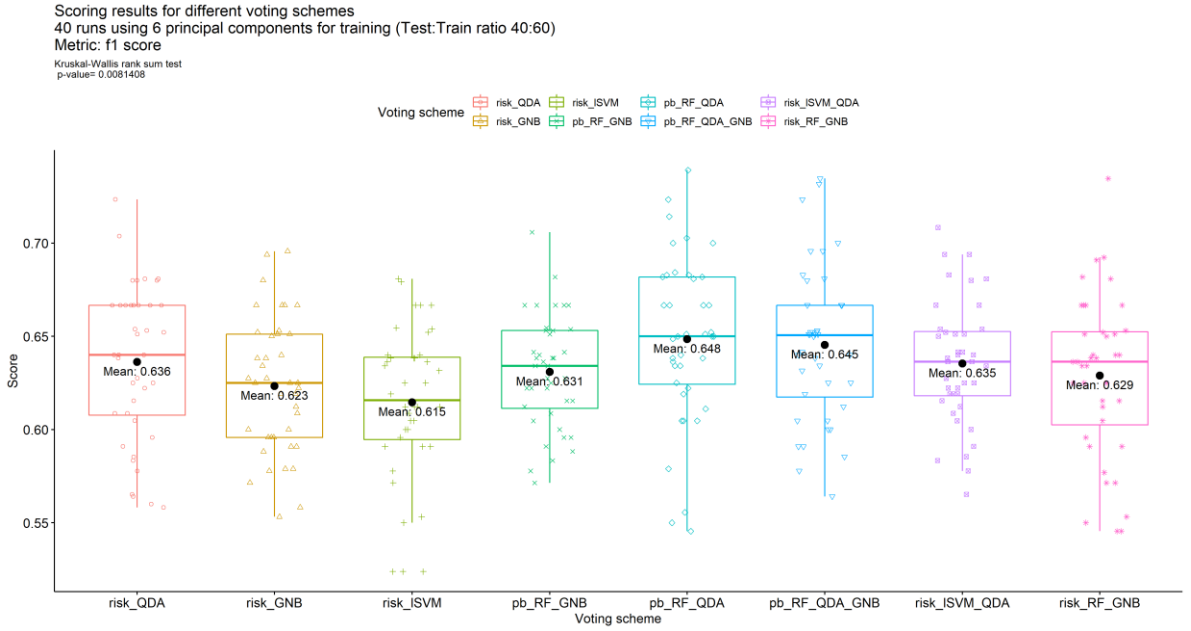


Figure 77: Optimal risk cut-off values in the 40 different train-test splits of the dataset, derived from survival analysis during the first 2.5 years. The 8 best performing, in terms of lowest standard deviation, classifiers and ensemble models (trained with 6 principal components and with a test ratio of 0.4) are being presented.

F1-score

Using mean as cut-off value for patients' risk stratification



Using median as cut-off value for patients' risk stratification

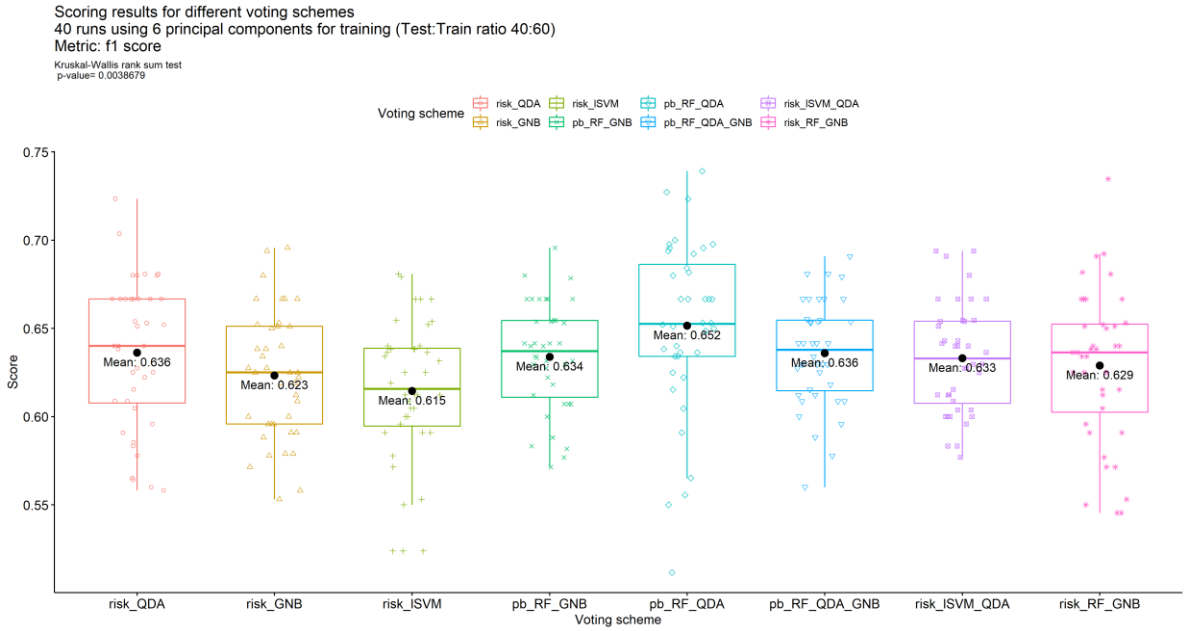
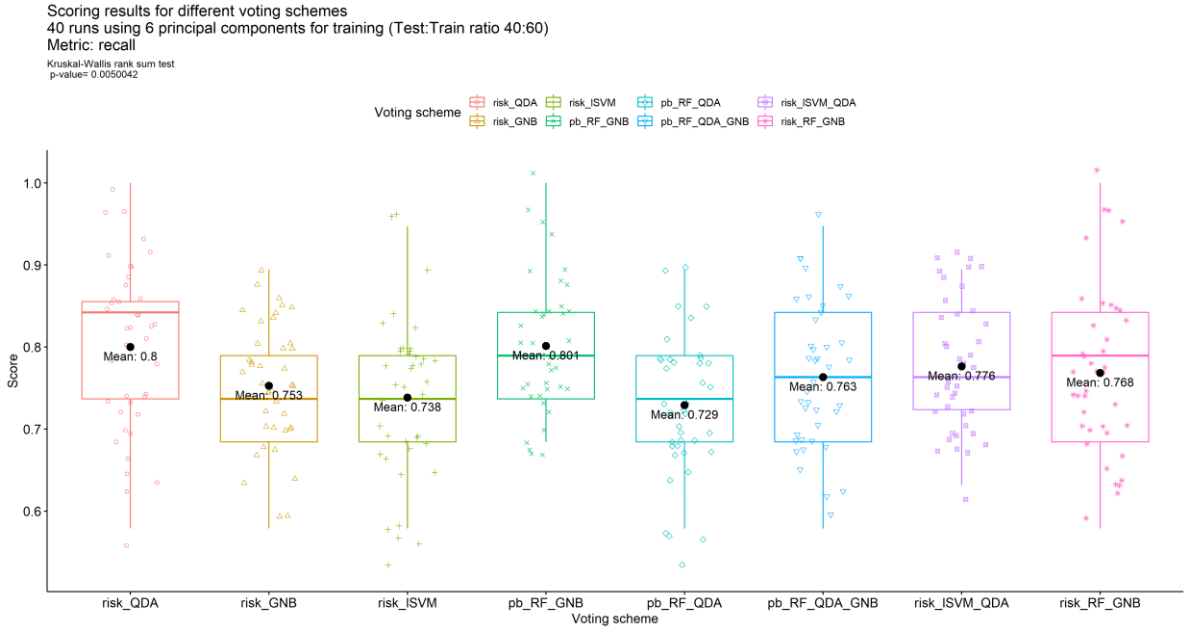


Figure 78: F1-score results in the validation set for classifiers and ensemble models trained with 6 principal components and with a test ratio of 0.4

Recall

Using mean as cut-off value for patients' risk stratification



Using median as cut-off value for patients' risk stratification

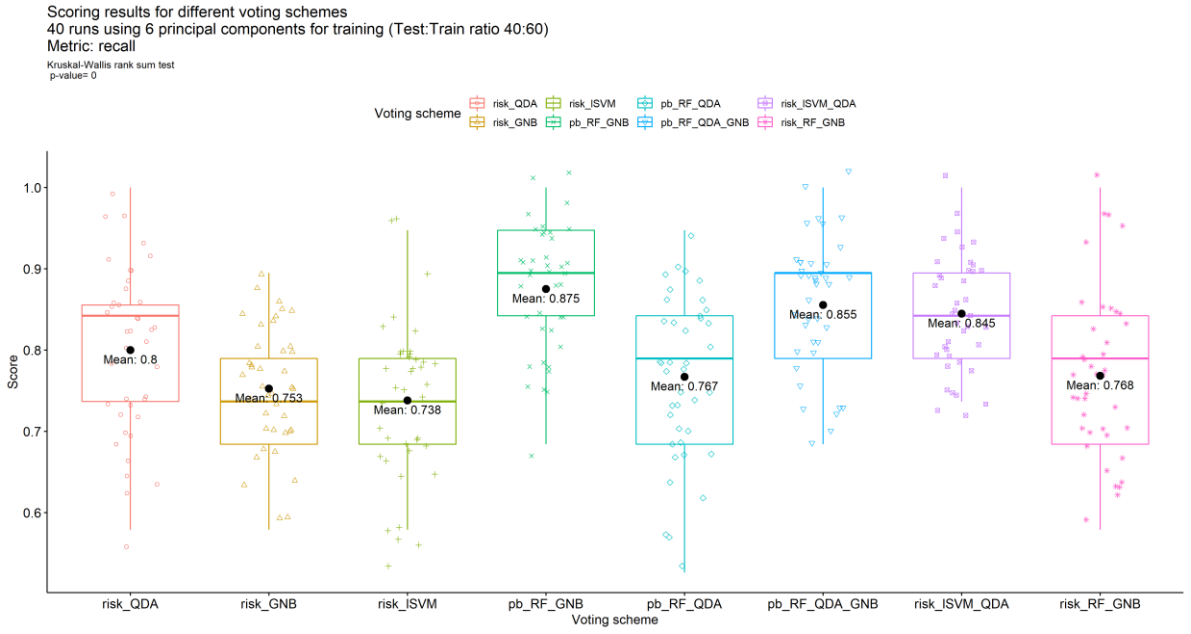


Figure 79: Recall results in the validation set for classifiers and ensemble models trained with 6 principal components and with a test ratio of 0.4

3.2.2.1.d Training with 7 principal components and with a test ratio of 0.4

GNB and lSVM classifiers achieved again the best results regarding MCC, balanced accuracy, ROC AUC score and bootstrapping (Figure 80), while hard voting based on GNB achieved better results in the validation set (Figures 81,82 and 83). Interestingly we observed that when the left-out for test ratio increases, hard voting models tend to have more stable results. Nevertheless, the better performing classifiers remain the same in

all cases, GNB and lSVM. This consistent finding is quite promising regarding the generalization of our results.

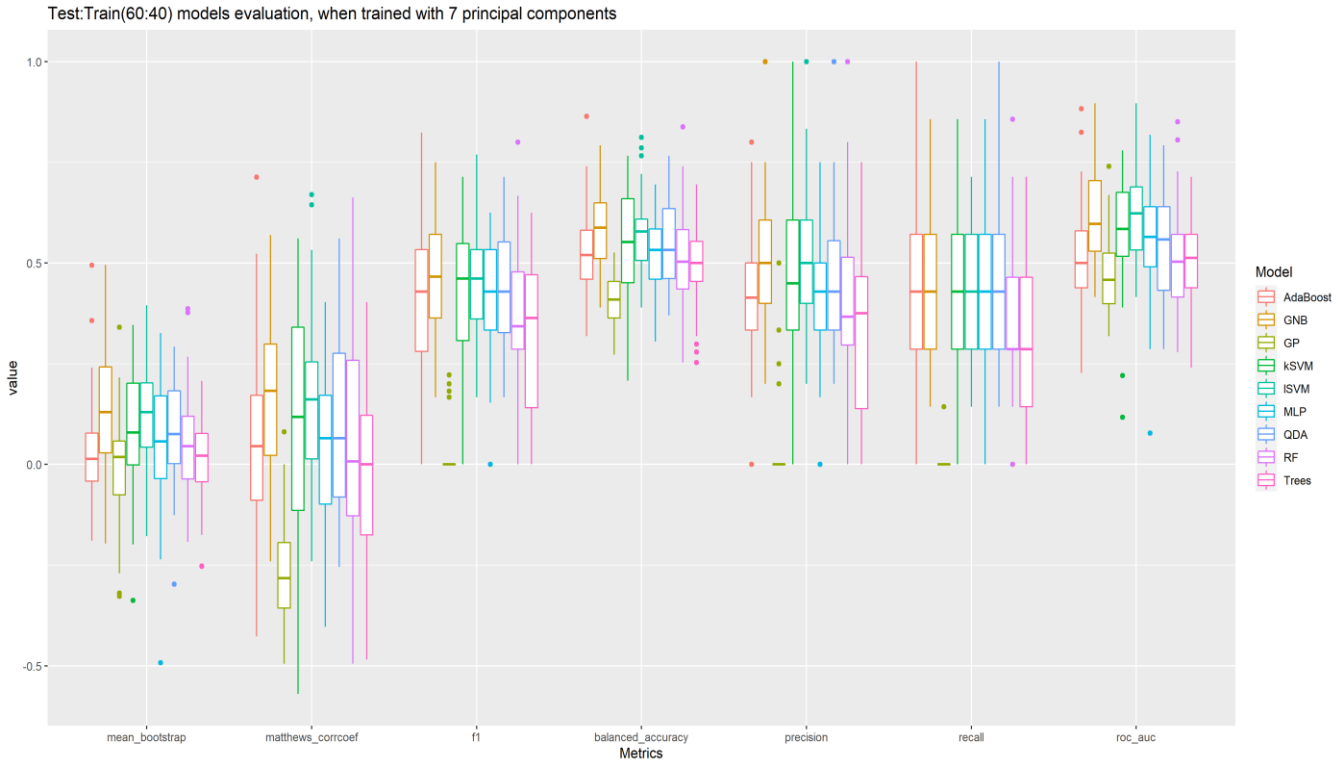


Figure 80: Training evaluation of different classifiers when trained with 7 principal components and with a test ratio of 0.4

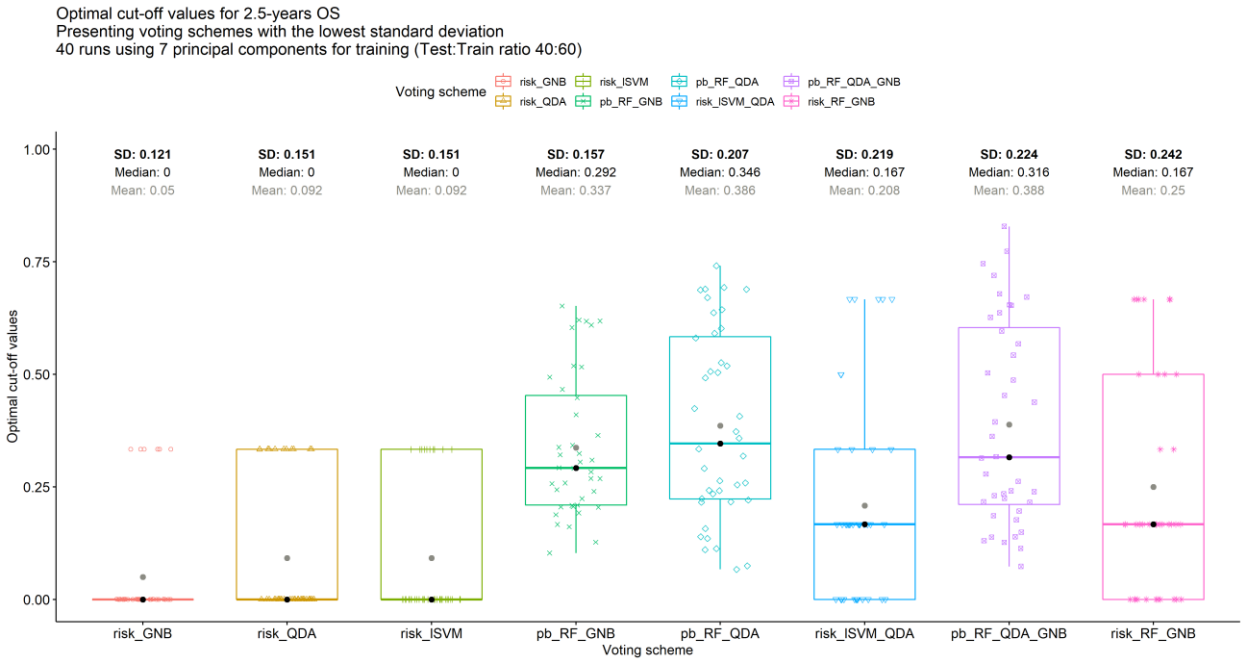
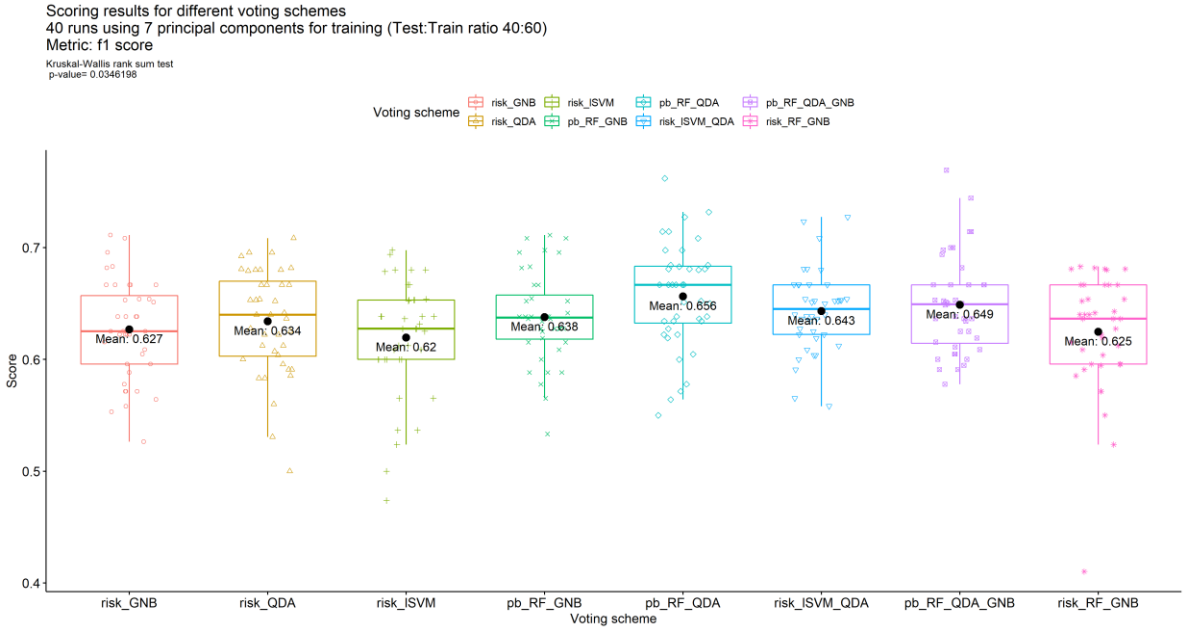


Figure 81: Optimal risk cut-off values in the 40 different train-test splits of the dataset, derived from survival analysis during the first 2.5 years. The 8 best performing, in terms of lowest standard deviation, classifiers and ensemble models (trained with 7 principal components and with a test ratio of 0.4) are being presented.

F1-score

Using mean as cut-off value for patients' risk stratification



Using median as cut-off value for patients' risk stratification

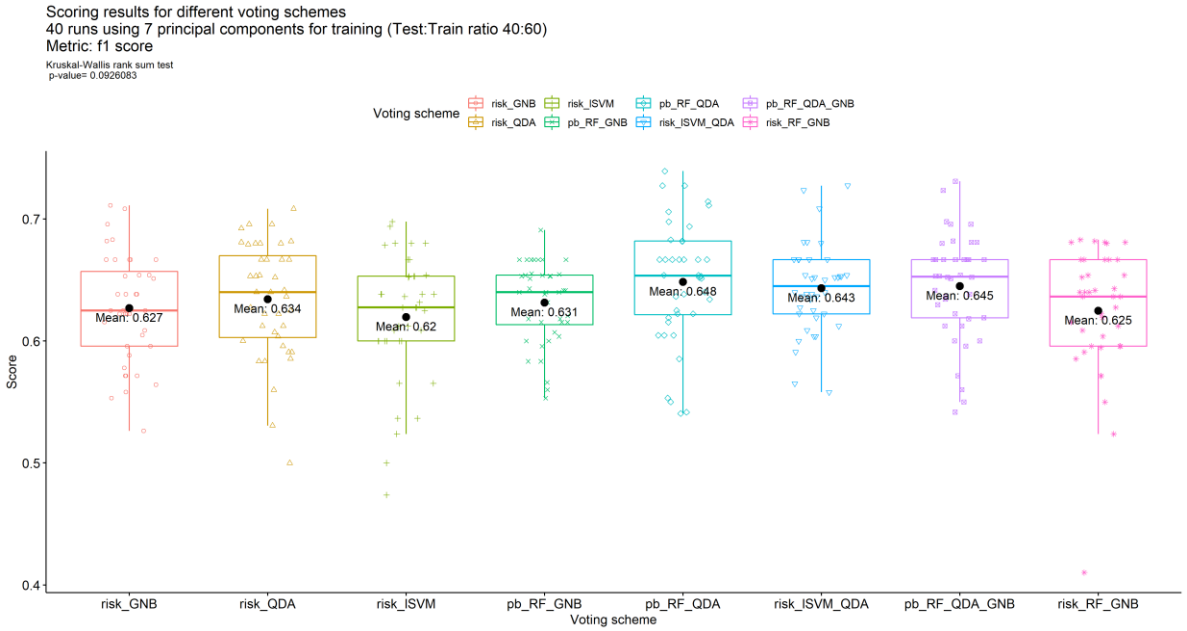
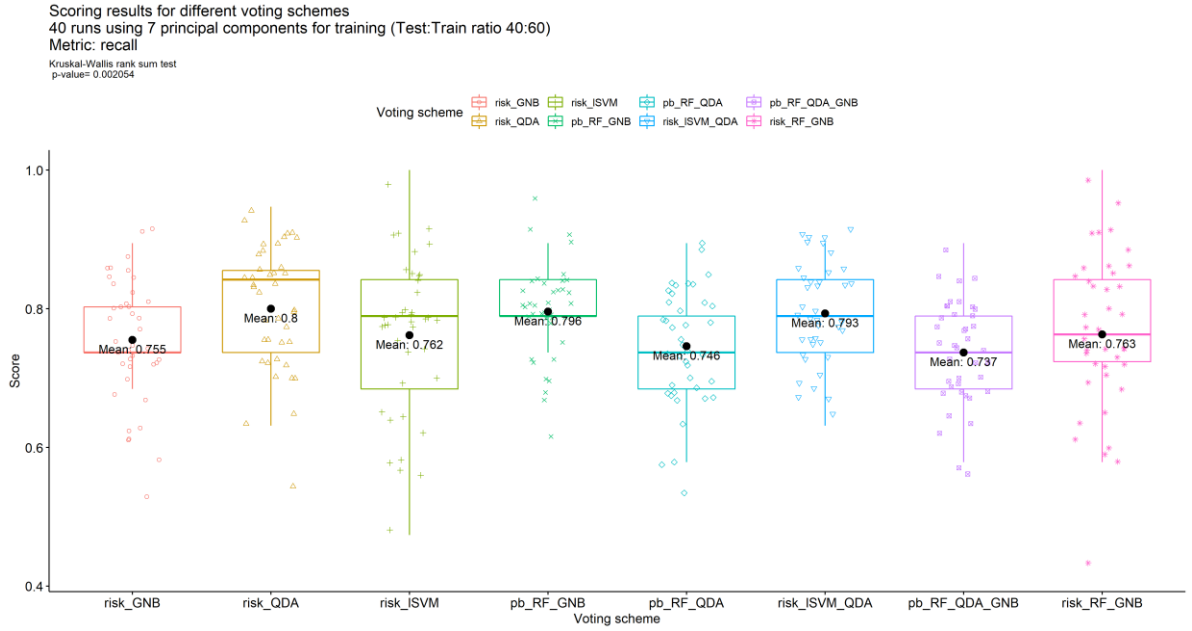


Figure 82: F1-score results in the validation set for classifiers and ensemble models trained with 7 principal components and with a test ratio of 0.4

Recall

Using mean as cut-off value for patients' risk stratification



Using median as cut-off value for patients' risk stratification

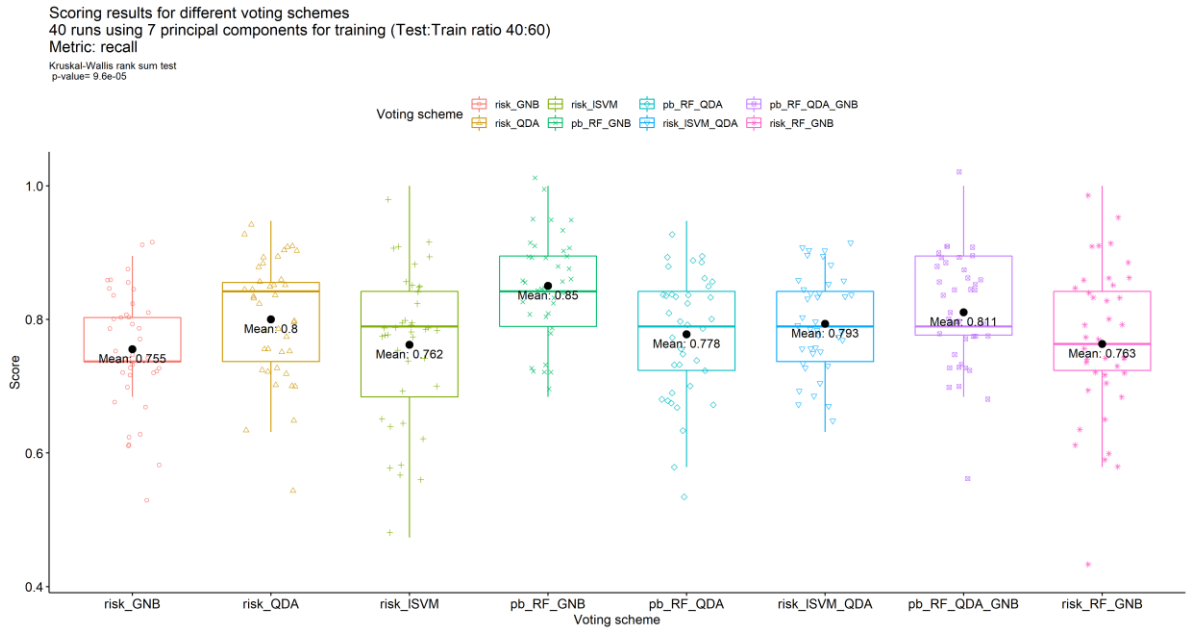


Figure 83: Recall results in the validation set for classifiers and ensemble models trained with 7 principal components and with a test ratio of 0.4

3.2.2.2 Supervised learning results on auto-segmented CT images - 51 patients

The inclusion of 4 extra HPV (-) OPC patients with OS >5 years resulted in a cohort with 51 patients (see Table 13) that remained to some extent balanced (ratio of premature death: 0.37). However, the overall results were inferior to those obtained in the more balanced cohort with 47 patients. Therefore, we will present only the results obtained with the usage of 7 principal components and with a test ratio of 0.3, a combination that have already been found to be the most promising.

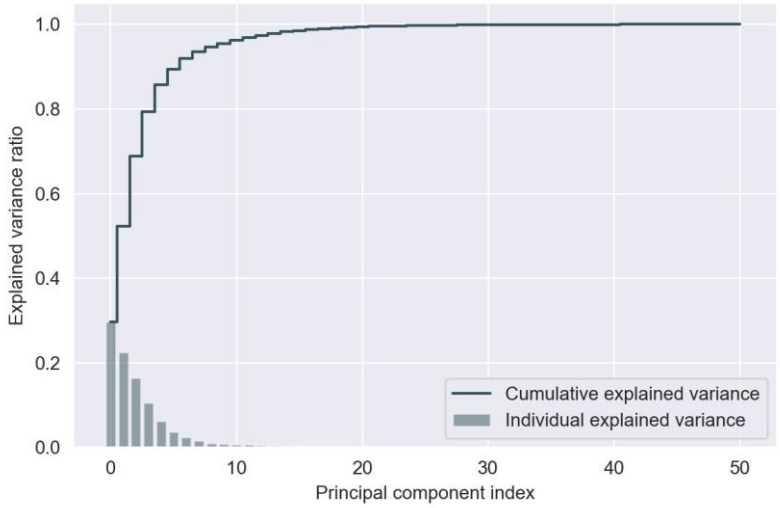


Figure 84: PCA analysis prior to training for the cohort with 51 patients

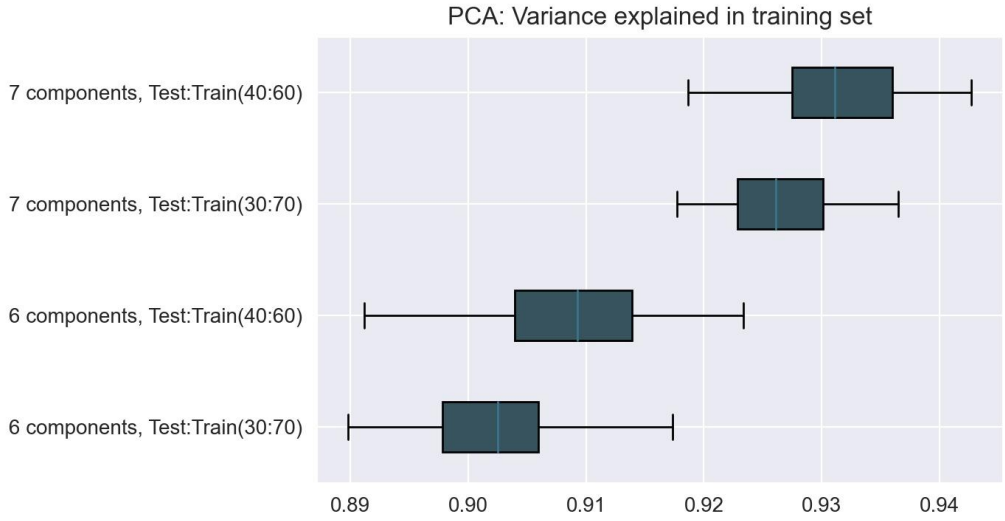


Figure 85: Variance explained in the training set by number of components and left out percentage for test, in the 40 different train-test splits of the dataset (cohort with 51 patients)

In consistent with the results in the cohort with 47 patients, GNB and lSVM classifiers achieved again the best training results regarding MCC, balanced accuracy and bootstrapping (Figure 86). In the validation set, hard voting of GNB and lSVM classifiers along with the soft voting ensemble models pb_RF_GNB and pb_RF_QDA_GNB achieved the best results (Figures 87, 88 and 89). In Figures 88 and 89 we presented only the results derived when using the median of all optimal cut-off values as cut-off point, based on previous findings that median leads to better recall results and consequently identification of more true high-risk patients. The resulting higher sensitivity when using median better serves the screening purposes of the proposed patients' risk stratification.

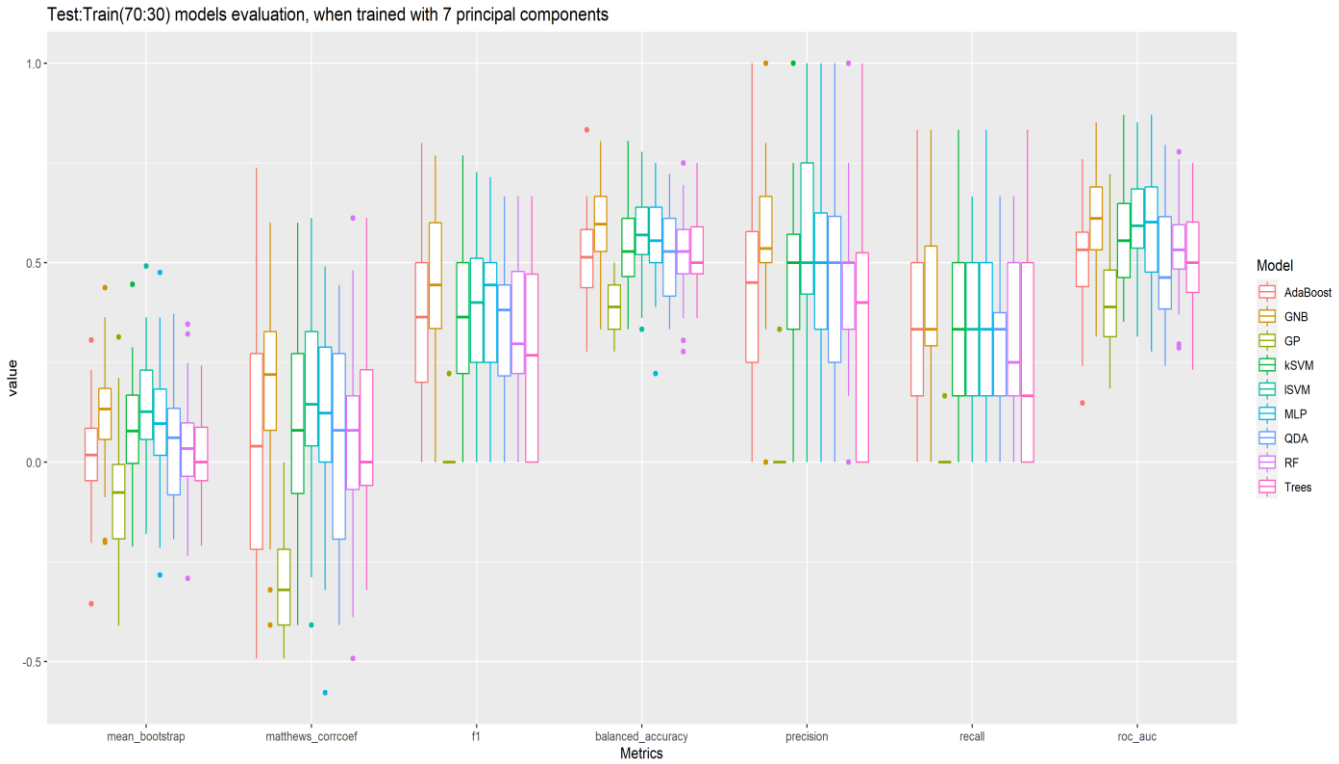


Figure 86: Training evaluation of different classifiers when trained with 7 principal components and with a test ratio of 0.3 (cohort with 51 patients)

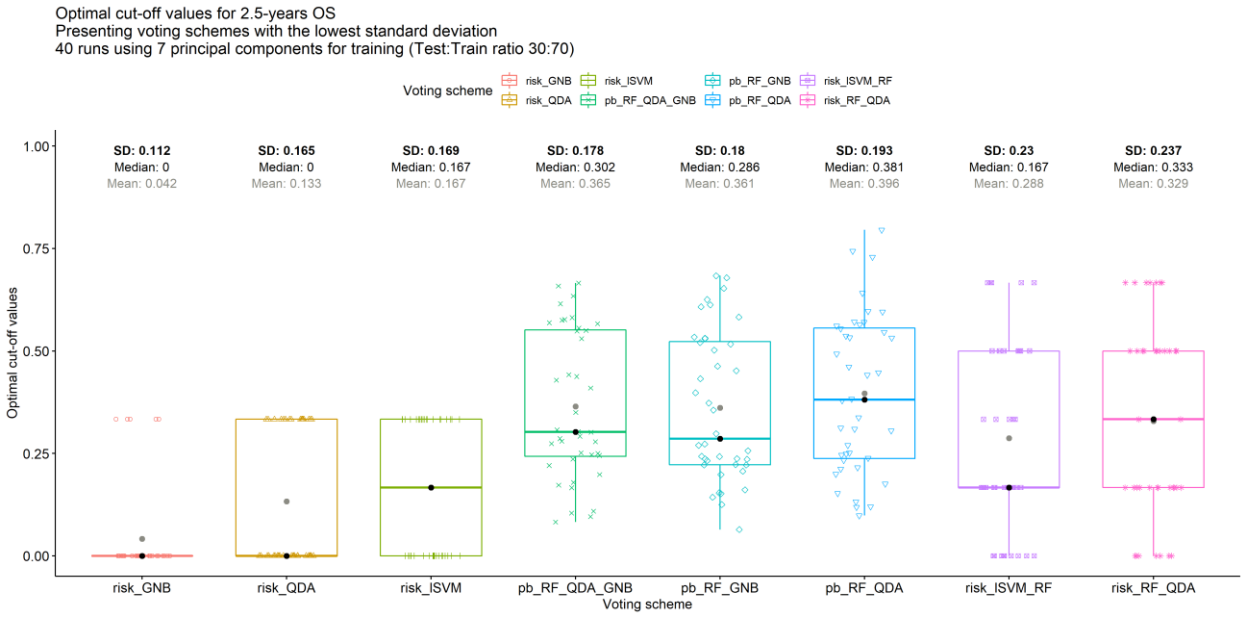


Figure 87: Optimal risk cut-off values in the 40 different train-test splits of the dataset, derived from survival analysis during the first 2.5 years. The 8 best performing, in terms of lowest standard deviation, classifiers and ensemble models (trained with 7 principal components and with a test ratio of 0.3) are being presented (cohort with 51 patients).

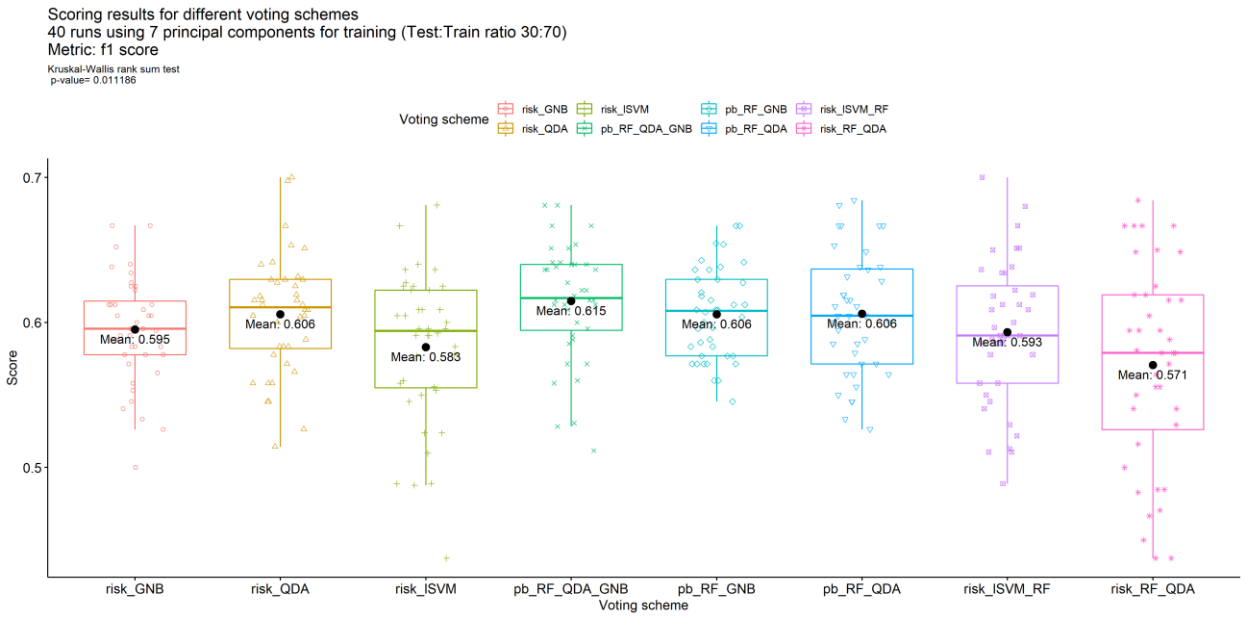


Figure 88: F1-score results in the validation set for classifiers and ensemble models trained with 7 principal components and with a test ratio of 0.3 (median used as cut-off, cohort with 51 patients)

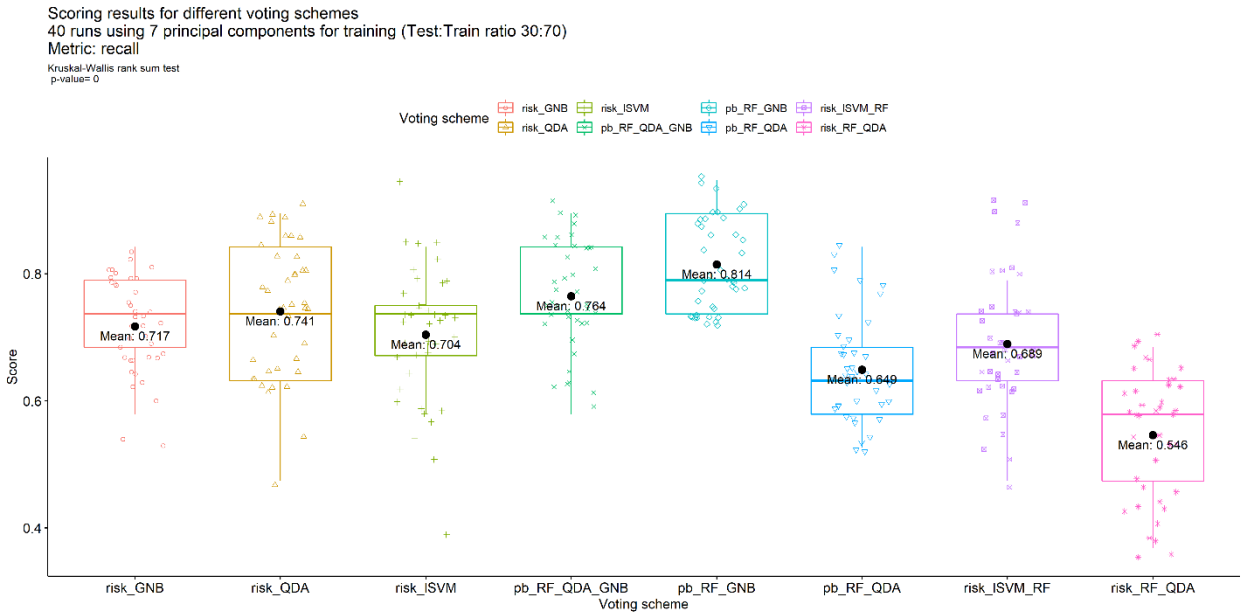


Figure 89: Recall results in the validation set for classifiers and ensemble models trained with 7 principal components and with a test ratio of 0.3 (median used as cut-off, cohort with 51 patients)

3.2.3 Survival results on auto-segmented CT images

For the classifiers with the better performance in terms of minimum optimal risk cut-offs' standard deviation , we used the same cut-off point for all the 40 different train-test splits of the dataset, in order to stratify patients in the validation set. As cut-off point, we tried both the mean and the median of all the optimal cut-off values. We then tested the percentage of the 40 different splits, where the proposed classification model achieved to separate survival of low-risk and high-risk group of patients statistically significantly (p-value < 0.05) or to showcase a trend for difference in survival between the two risk groups

(p-value < 0.1). By using the same cut-off, and not the optimal, we tried to partially overcome the bias of an outcome-oriented method (providing a value of a cut-point that correspond to the most significant relation with outcome, in our case survival) and to present results with some potential of generalisation. Our survival results confirmed that in most cases the resulting higher sensitivity when using median better serves the screening purposes of the proposed patients' risk stratification. Moreover, we also confirmed that models with better performance regarding both F1-score and recall were the ones that led to better separation of the 2.5 years overall survival curves between patients classified as high and low risk. In subsections 3.2.3.1 and 3.2.3.2 we will present classification models' results obtained when using the median as cut-off point. We selected to present the best results obtained in different training settings (7 principal components with a test ratio of 0.3, 6 principal components with a test ratio of 0.3 and 7 principal components with a test ratio of 0.4).

3.2.3.1 Survival results on auto-segmented CT images - 47 patients

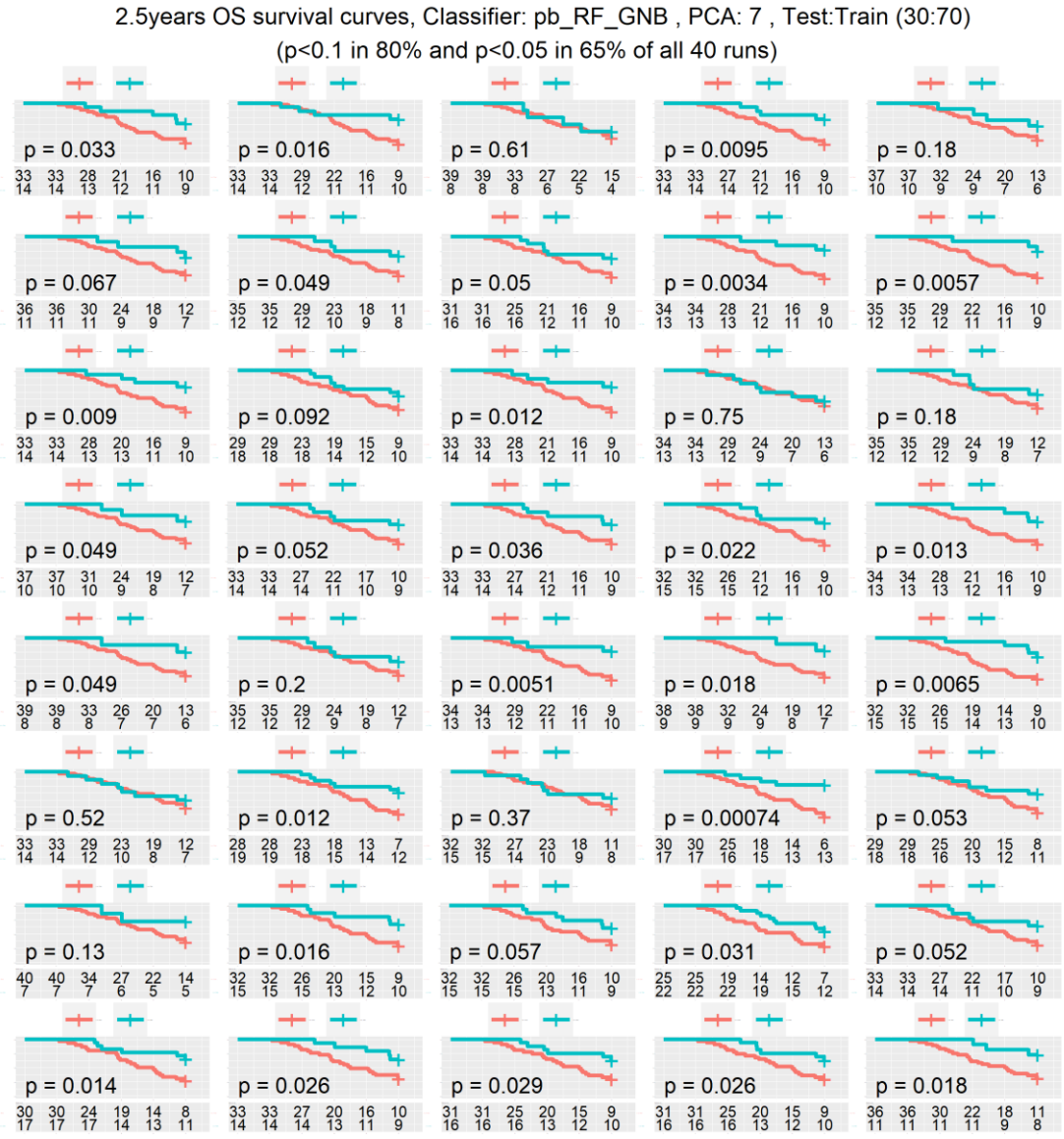


Figure 90: Kaplan–Meier 2.5 years overall survival curves according to risk group classification in the validation set (red curve for the high-risk group and blue curve for the low-risk group), by the pb_RF_GNB classifiers (trained with 7 principal components and with a test ratio of 0.3 in the 40 different train-test splits of the dataset), using in all cases the same risk cut-off value (the median of all 40 optimal-cut-off values); cohort with 47 patients

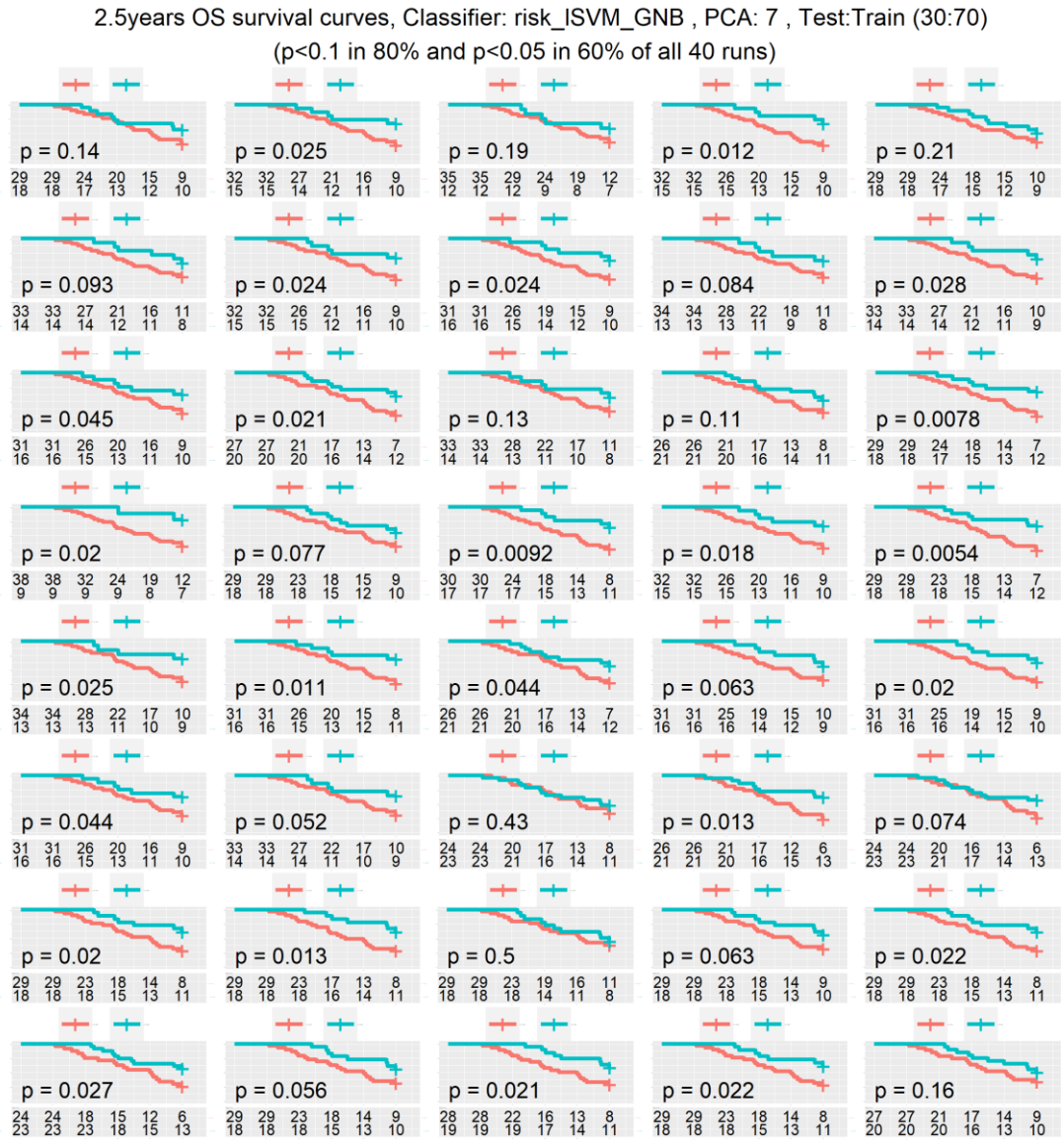


Figure 91: Kaplan–Meier 2.5 years overall survival curves according to risk group classification in the validation set (red curve for the high-risk group and blue curve for the low-risk group), by the risk_lSVM_GNB classifiers (trained with 7 principal components and with a test ratio of 0.3 in the 40 different train-test splits of the dataset), using in all cases the same risk cut-off value (the median of all 40 optimal-cut-off values); cohort with 47 patients

The best results in the validation set were obtained in the cohort with 47 patients and when classification models were trained with 7 principal components and with a test ratio of 0.3 (Figures 90, 91) followed by the results obtained when trained with 6 principal components and with a test ratio of 0.3 (Figures 92, 93).

In the case of training with 7 principal components the best performing classifiers were the soft voting model pb_RF_GNB and the hard voting model risk_lSVM_GNB. Both models achieved to showcase a trend for difference in survival between the two risk groups (p-value < 0.1) in 80% of the 40 different train-test splits of the dataset.

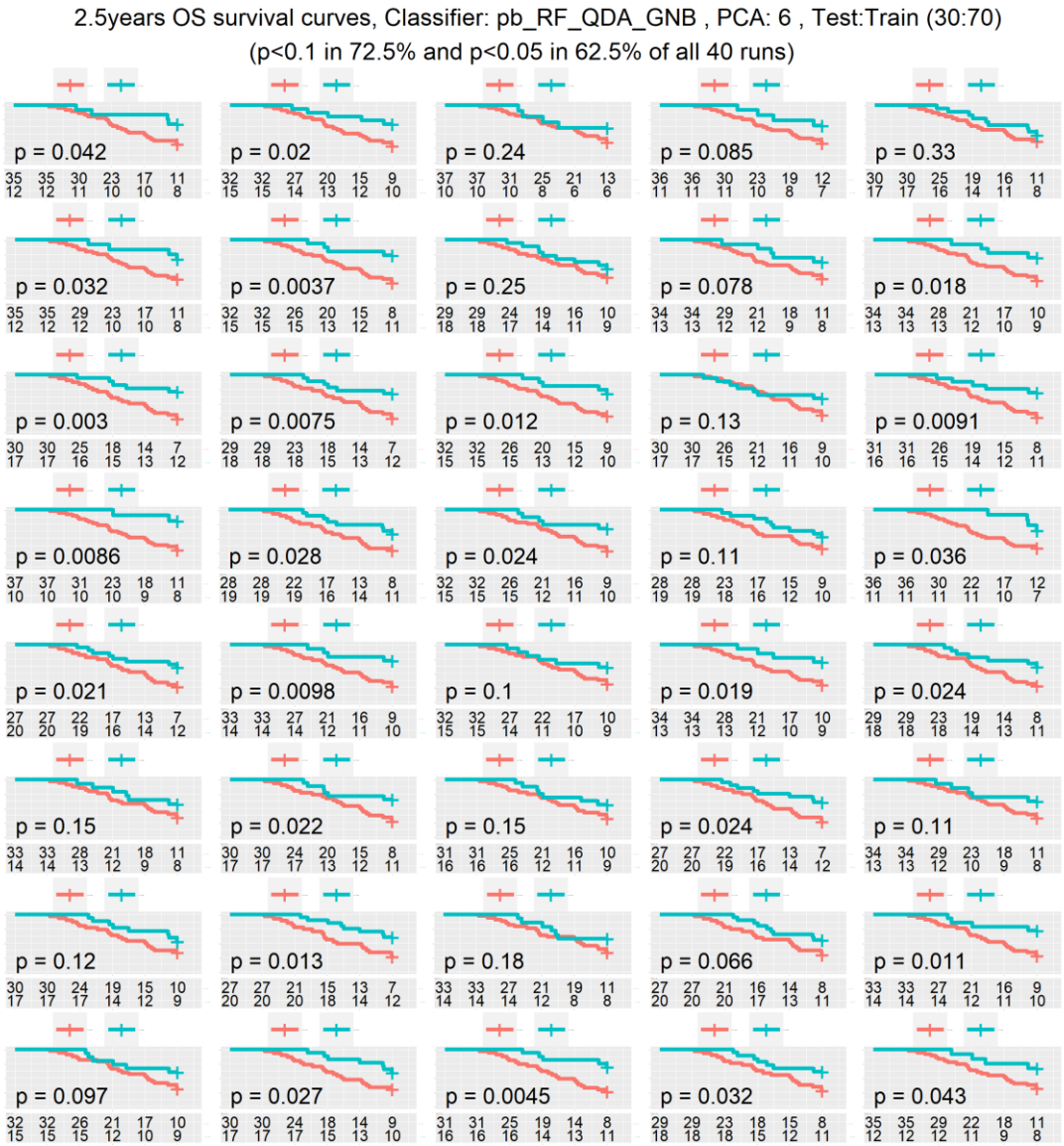


Figure 92: Kaplan–Meier 2.5 years overall survival curves according to risk group classification in the validation set (red curve for the high-risk group and blue curve for the low-risk group), by the pb_RF_QDA_GNB classifiers (trained with 6 principal components and with a test ratio of 0.3 in the 40 different train-test splits of the dataset), using in all cases the same risk cut-off value (the median of all 40 optimal-cut-off values); cohort with 47 patients

In the case of training with 6 principal components the best performing classifiers were the soft voting model pb_RF_QDA_GNB and the hard voting model risk_GNB_QDA_RF_lSVM. Both models achieved to showcase a trend for difference in survival between the two risk groups (p-value < 0.1) in around 70% of all the 40 different train-test splits of the dataset. Notably, the hard voting model risk_GNB_QDA_RF_lSVM achieved the best overall results regarding separating survival curves of low-risk and high-risk group of patients statistically significantly (p-value < 0.05), reaching such results in 67.5% of all the 40 different train-test splits of the dataset (Figure 93).

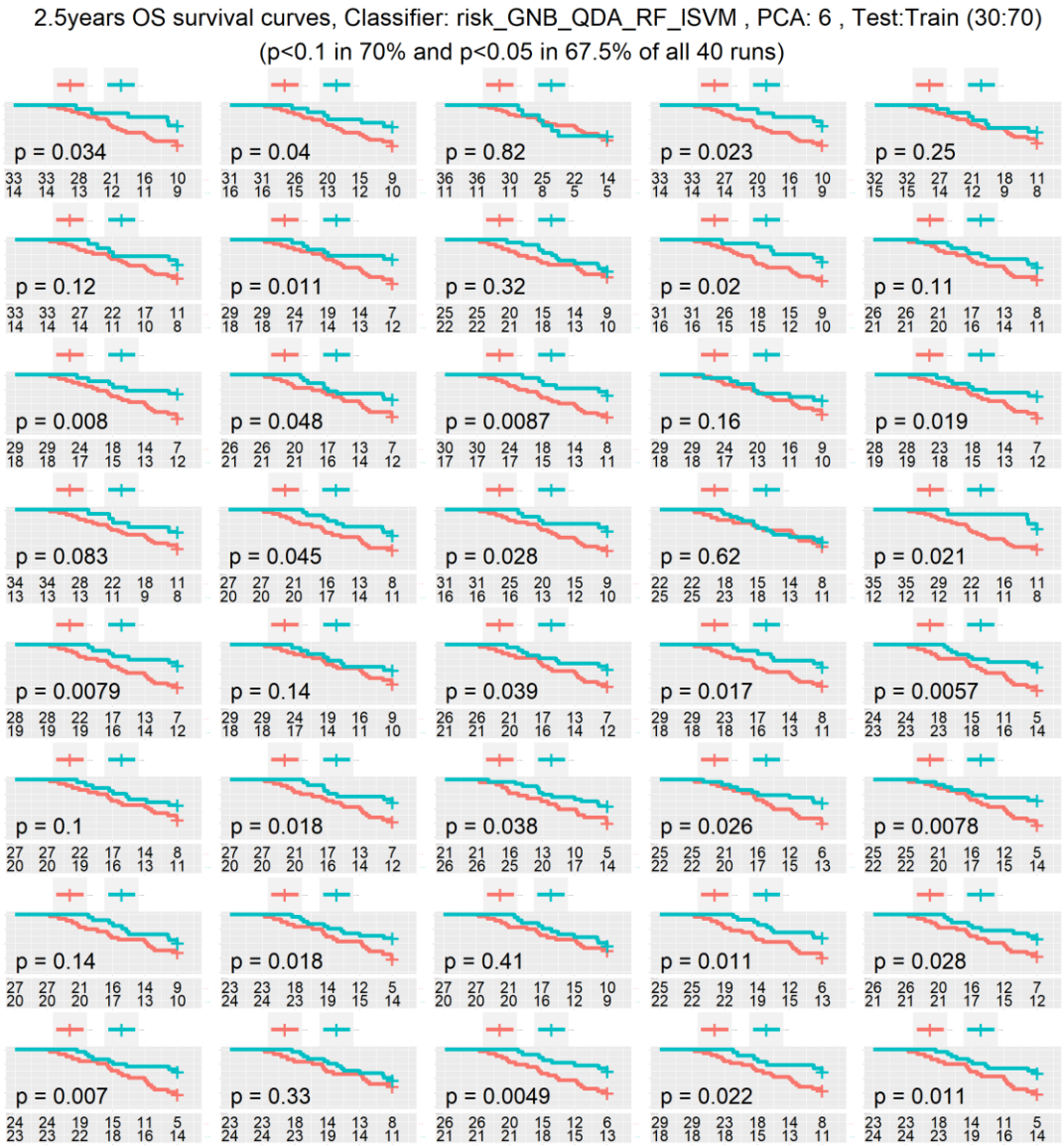


Figure 93: Kaplan–Meier 2.5 years overall survival curves according to risk group classification in the validation set (red curve for the high-risk group and blue curve for the low-risk group), by the risk_GNB_QDA_RF_lSVM classifiers (trained with 6 principal components and with a test ratio of 0.3 in the 40 different train-test splits of the dataset), using in all cases the same risk cut-off value (the median of all 40 optimal-cut-off values); cohort with 47 patients

Finally, even in the case of training with 7 principal components and with the quite high test ratio of 0.4, hard voting with GNB alone achieved in the validation set to showcase a trend for difference in survival between the two risk groups (p-value < 0.1) in 70% of all the 40 different train-test splits of the dataset (Figure 94). Looking back at Figure 81 we can see that the cut-off point used for the risk_GNB was 0, meaning that in the validation set only one out of the three images per patient had to be classified as high-risk, by the GNB classifier, in order to finally classify the patient as high-risk. However, the results were quite inferior in separating survival curves of low-risk and high-risk group of patients statistically significantly (p-value < 0.05), achieving such results in only 45% of the 40 different train-test splits of the dataset (Figure 94).

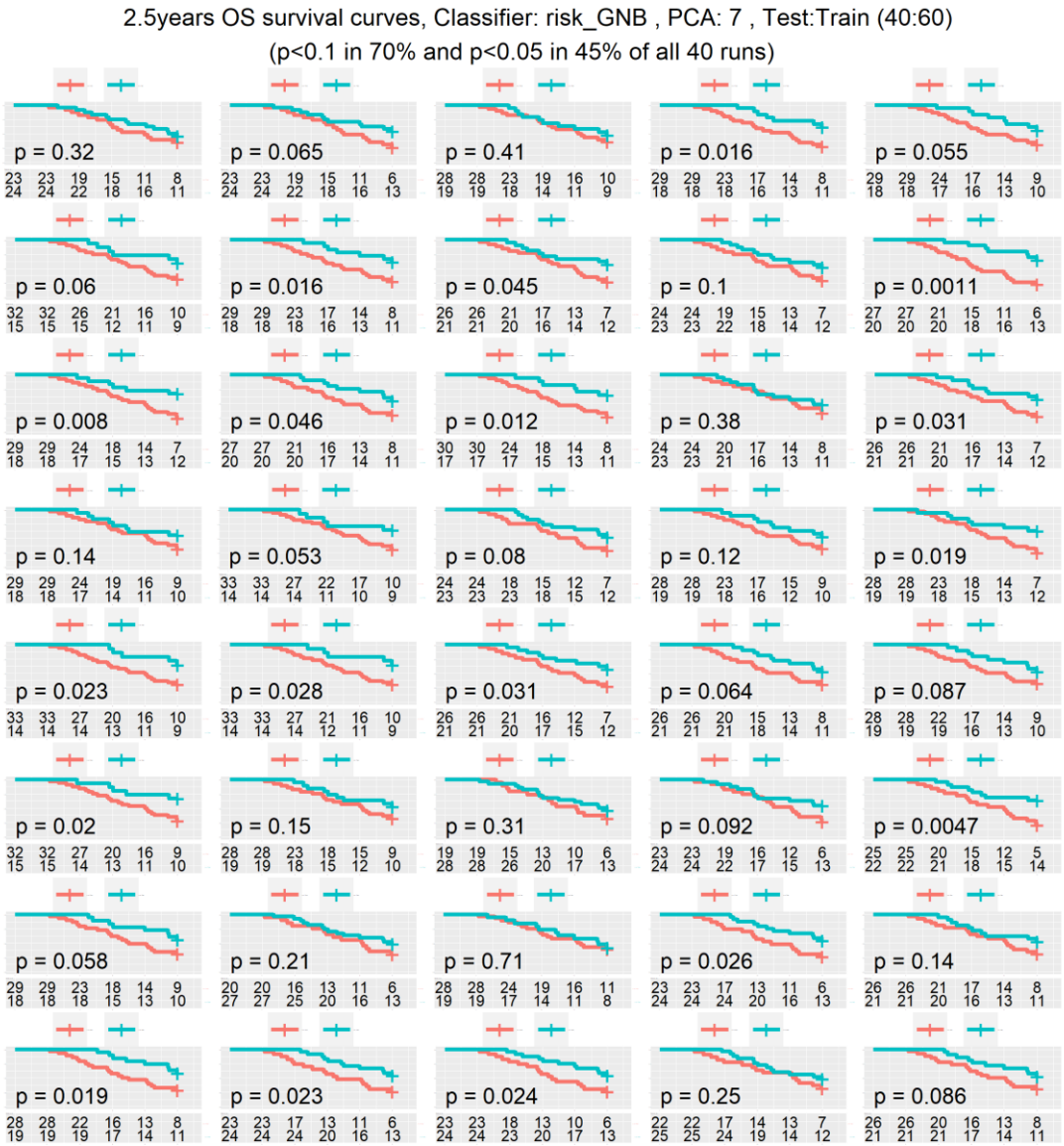


Figure 94: Kaplan–Meier 2.5 years overall survival curves according to risk group classification in the validation set (red curve for the high-risk group and blue curve for the low-risk group), by the risk_GNB classifiers (trained with 7 principal components and with a test ratio of 0.4 in the 40 different train-test splits of the dataset), using in all cases the same risk cut-off value (the median of all 40 optimal-cut-off values); cohort with 47 patients

3.2.3.2 Survival results on auto-segmented CT images – 51 patients

In the cohort with 51 patients the results were inferior to those of the cohort with 47 patients. Only classifiers resulted from the training setting with 7 principal components and with a test ratio of 0.3 achieved to showcase adequate separation between the survival curves of the low and high risk group of patients. Again, the GNB classifiers achieved the best results followed by the lSVM classifiers and the ensemble soft voting model pb_RF_GNB (Figures 95, 96 and 97).

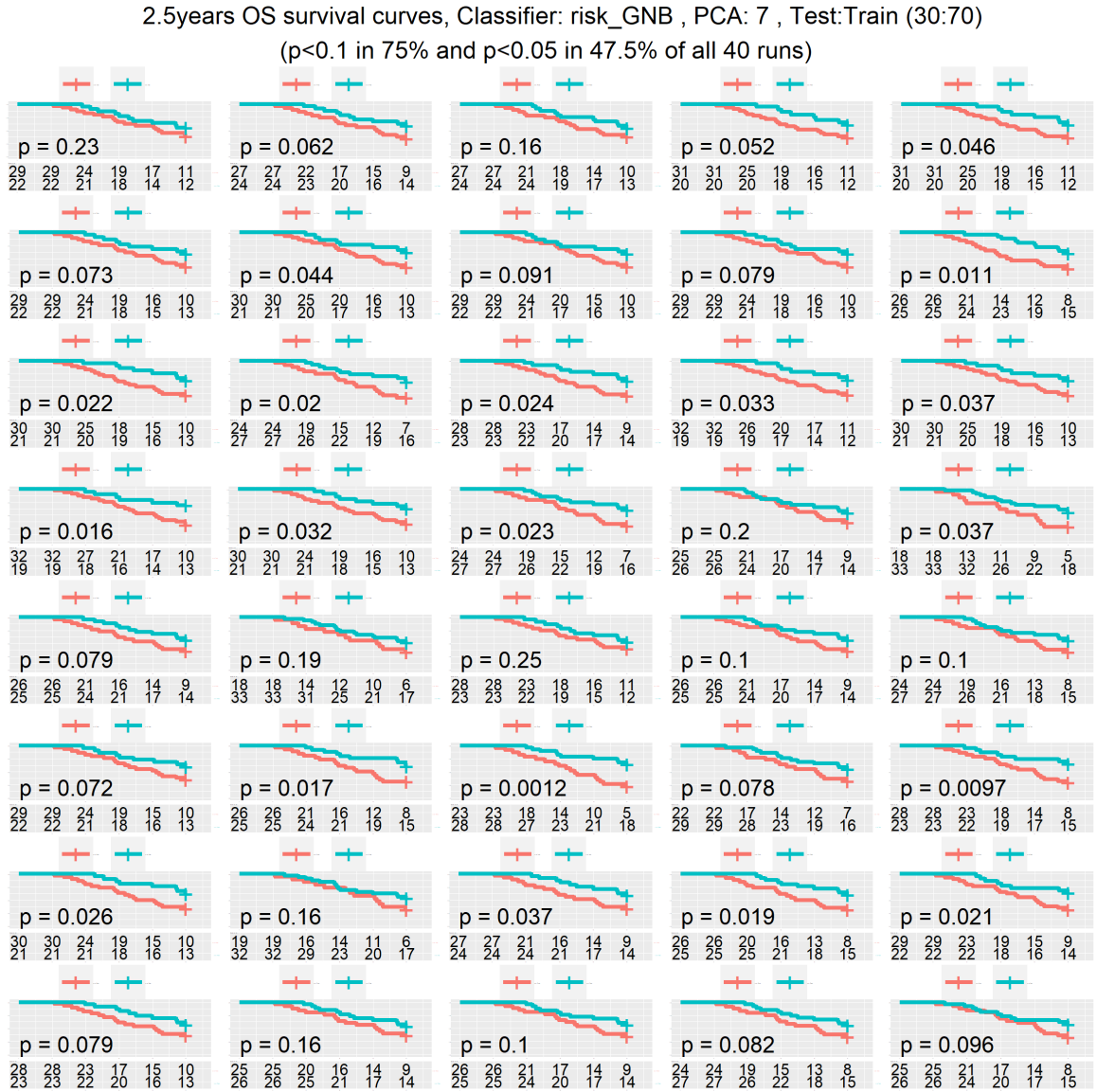


Figure 95: Kaplan–Meier 2.5 years overall survival curves according to risk group classification in the validation set (red curve for the high-risk group and blue curve for the low-risk group), by the risk_GNB classifiers (trained with 7 principal components and with a test ratio of 0.3 in the 40 different train-test splits of the dataset), using in all cases the same risk cut-off value (the median of all 40 optimal-cut-off values); cohort with 51 patients

Although those classification models achieved to showcase a trend for difference in survival between the two risk groups (p-value < 0.1), in high percentages of the 40 different train-test splits of the dataset (reaching even 75%), those models had poorer results in separating survival curves of the low-risk and the high-risk group of patients statistically significantly (p-value < 0.05), achieving such results in only 47.5%-55% of the 40 different train-test splits of the dataset (Figures 95, 96 and 97).

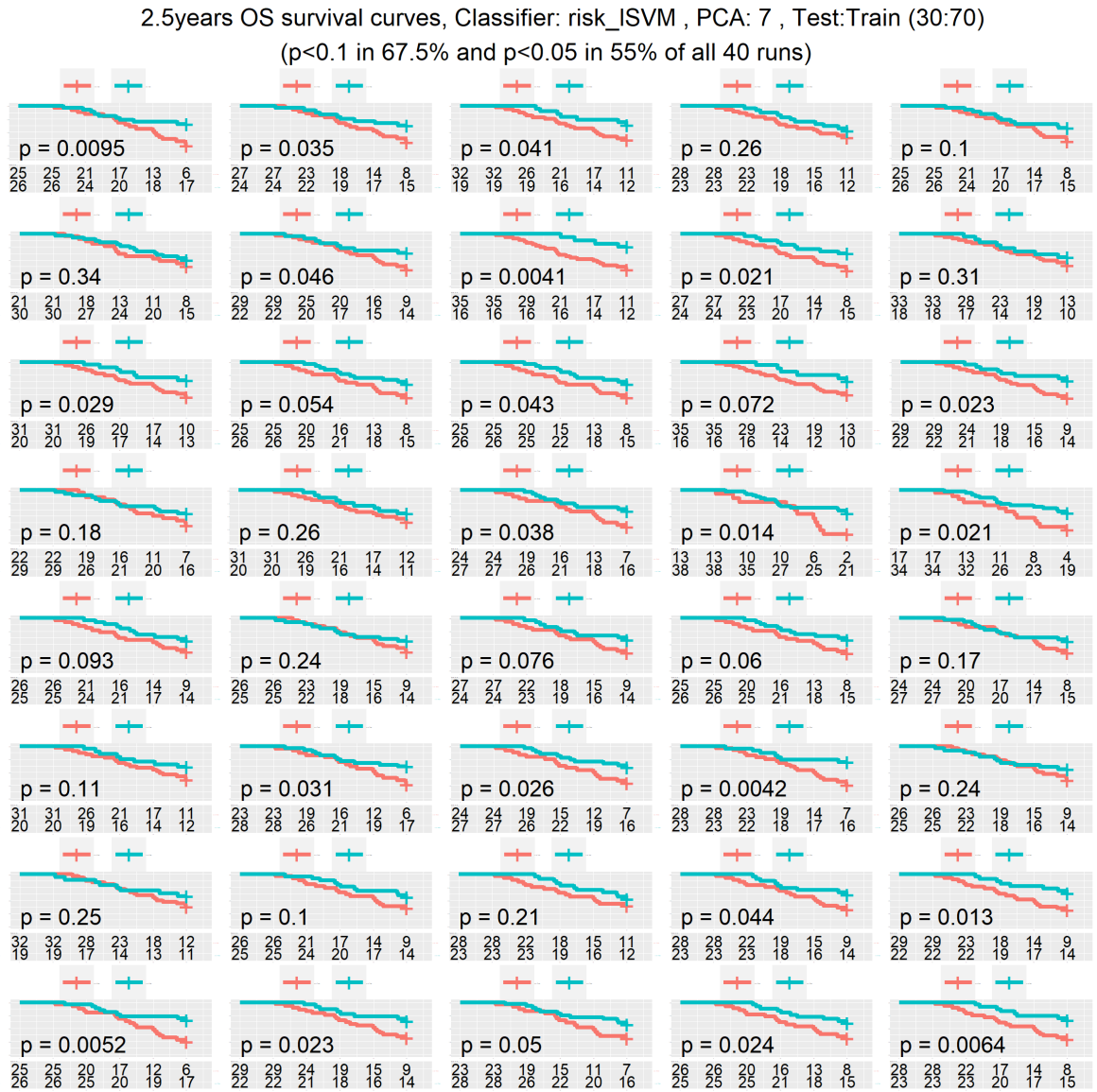


Figure 96: Kaplan–Meier 2.5 years overall survival curves according to risk group classification in the validation set (red curve for the high-risk group and blue curve for the low-risk group), by the risk_lSVM classifiers (trained with 7 principal components and with a test ratio of 0.3 in the 40 different train-test splits of the dataset), using in all cases the same risk cut-off value (the median of all 40 optimal-cut-off values); cohort with 51 patients

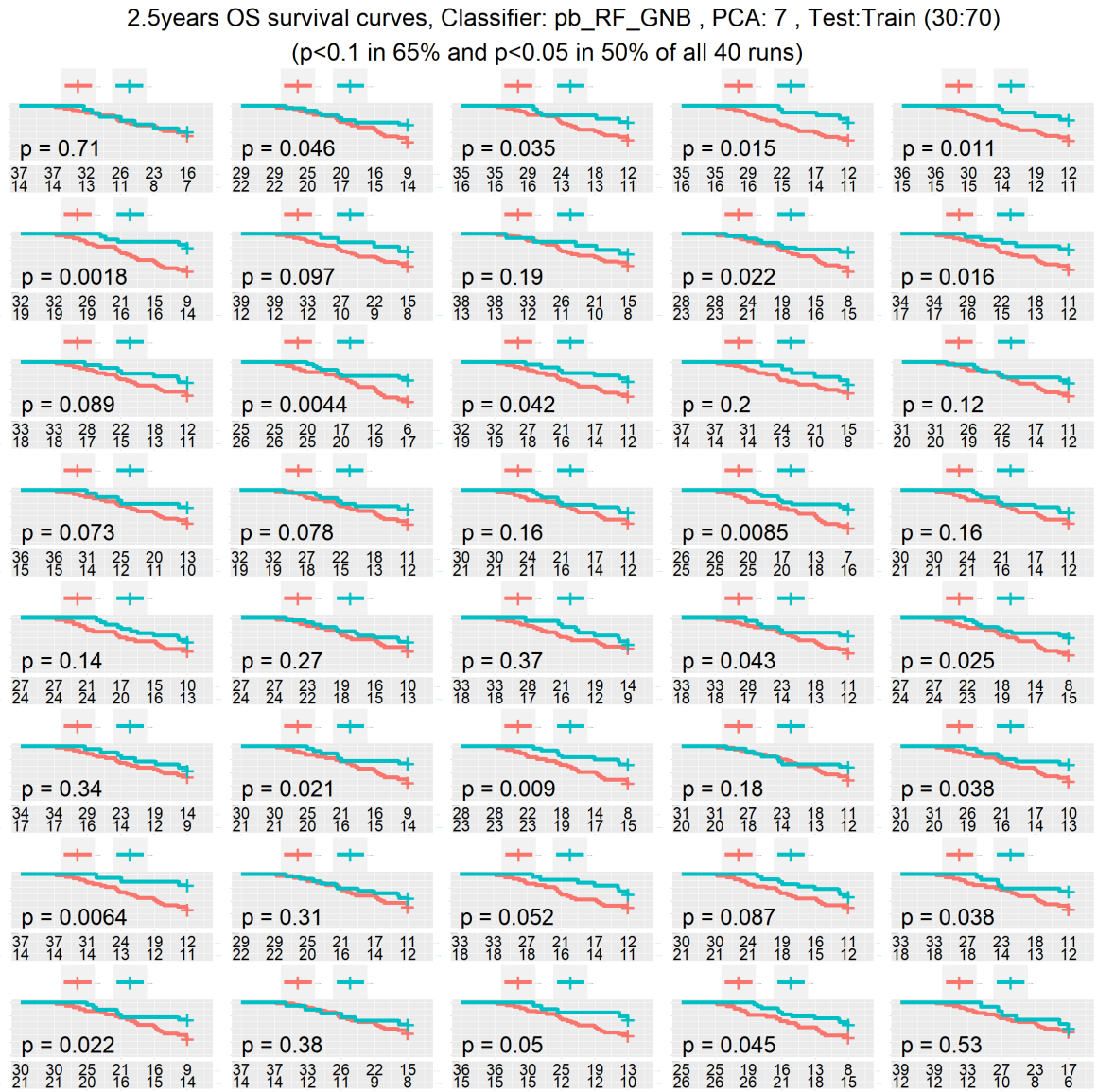


Figure 97: Kaplan–Meier 2.5 years overall survival curves according to risk group classification in the validation set (red curve for the high-risk group and blue curve for the low-risk group), by the pb_RF_GNB classifiers (trained with 7 principal components and with a test ratio of 0.3 in the 40 different train-test splits of the dataset), using in all cases the same risk cut-off value (the median of all 40 optimal-cut-off values); cohort with 51 patients

4. DISCUSION

Undefined HPV status in oropharyngeal carcinomas was a serious limitation for the current pilot study. The unknown HPV status acts as a highly confounding factor. With the application of further exclusion criteria in the group of patients with oropharyngeal carcinomas we partially overcame it. On the one hand, by excluding all patients with oropharyngeal carcinomas and overall survival greater than 5 years, a group of patients that was considered as a potential confounder in our study, we ended up with a balanced cohort in terms of survival categories. The most balanced cohort was the one with 47 patients, and our results showed that it favoured training, as we achieved the best results in that case. On the other hand, the exclusion of all these patients resulted in a small dataset.

Radiomics machine learning studies with small sample size are quite challenging and can lead to unreliable results. We applied unsupervised learning to investigate the separability of our data. We observed that when excluding patients with OPSCC without premature death, there seemed to be an inherent 3-cluster tendency in our dataset (one cluster with overrepresentation of low-risk patients and two clusters with overrepresentation of high-risk patients). We also observed that when taking all features into consideration there seem to be a stronger inherent tendency for clusters formation that were relevant to the clinical problem addressed. The aforementioned observations indicated that proceeding with the smaller but more balanced cohorts and with a dimensionality reduction method that takes all features and consequently the interaction effects among the features into consideration, like PCA, would be a promising strategy.

Our classification results were very encouraging, as we managed to train classifiers that served well the screening purposes of the problem addressed, by achieving high recall while maintaining an acceptable F1-score. The results were validated by survival analysis. By using the same cut-off, and not the optimal, we tried to partially overcome the bias of an outcome-oriented method and to present results with some potential of generalisation. It was also encouraging that in different training settings the same classifiers, GNB and lSVM, emerged as the ones with the better performance. Moreover, the soft voting ensemble model pb_RF_GNB was also consistently among the better performing.

Nevertheless, more data, with known HPV status for the OPSCC patients, are needed to achieve better and more stable results. Data augmentation by using more CT slices at the same cervical level (C3) per patient is not recommended especially as long as the dataset remains relatively small. However, using more than one image per patient might be beneficial when calculating the patient's risk for premature death and is recommended both for validation purposes on unseen data, and for classification of new entries whose outcome is yet unknown.

To conclude, prognosis of HNSCC patients remains complex and such risk classifications should be considered only in more complex models along with other well studied prognostic features. When other indications of malnutrition are present , nutritional interventions should be seriously considered in patients classified as high-risk. Sex and biometric measurements should be also taken into consideration in larger datasets. When sarcopenia can be defined from the L3 level, prevalence of sarcopenia in the high-risk group should also be addressed. Still, such studies cannot be conducted without the

establishment of large national cancer databases, and it is in this direction where we should focus our actions. Figure 98 shows how a head and neck cancer registry could be developed and utilized.

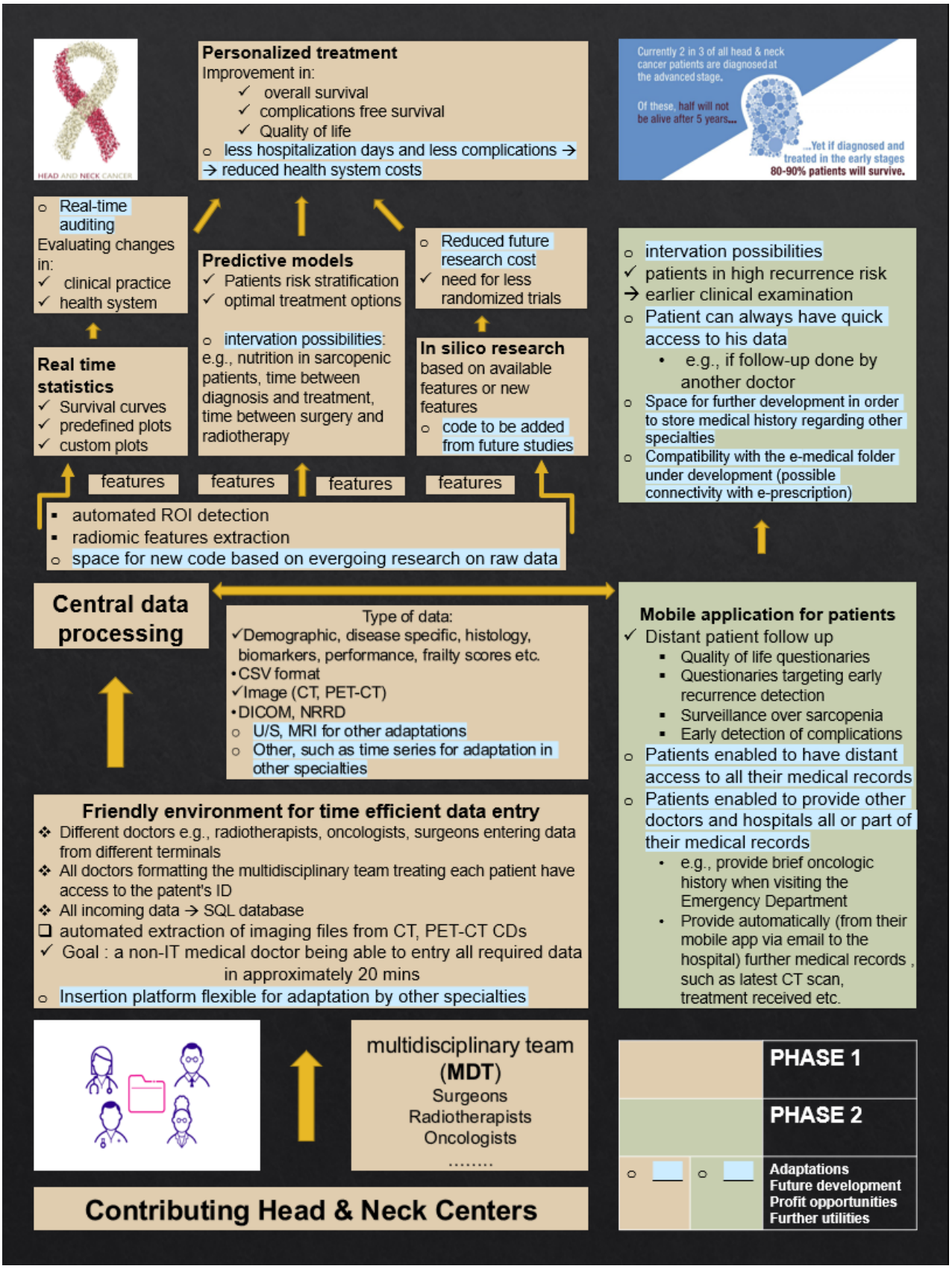


Figure 98: Head and neck cancer registry - data collection, processing, prediction models and future perspectives

5. CONCLUSION

The proposed automatic method for segmentation, radiomic feature extraction and subsequent patient risk stratification, based on CT-derived skeletal muscle related data, constitutes a promising automatic screening method. The fact that results were evaluated on 40 different train-test splits of the dataset and that proposed risk stratification was tested on a validation set using the same risk cut-off points and not always the optimal ones, along with the consistency regarding various classifiers' performance pave the way for potential generalization. However, more data are needed, with known HPV status for the OPSCC patients in order to establish risk stratification based on CT-derived skeletal muscle related data as a clinically useful biomarker, that might be integrated in more complex machine learning prognostic models aiming personalized treatment.

ABBREVIATIONS

AI, artificial intelligence

ASM, appendicular skeletal muscle mass

AUC-ROC, area under receiver operating characteristics curve

C3, third cervical spine vertebra

CC, consensus clustering

CGA, comprehensive geriatric assessment

CI, confidence interval

CSA, cross-sectional area

CSS, cancer specific survival

CT, computed tomography

DFS, disease free survival

EM, expectation-maximization

FI, frailty index

GAN, generative adversarial network

GLCM, gray-level co-occurrence matrix

GLDZM, gray-level distance-zone matrix

GLM, generalized linear model

GLRLM, gray-level run-length matrix

GLSZM, gray-level size-zone matrix

GMM, gaussian mixture models

GNB, gaussian naive Bayes

GP, gaussian processes,

HNC, head and neck cancer

HNSCC, head and neck squamous cell carcinoma

HPV, human papilloma virus

HR, hazard ratio

HU, Hounsfield unit

ICA, independent component analysis

ICC, intraclass correlation coefficient

IH, intensity histogram

IVH, intensity-volume histogram

KNN, k-nearest neighbour

kSVM, kernel support vector machine

L3, third lumbar spine vertebra

lSVM, linear support vector machine

MACT, mean muscle attenuation on CT scan

MCC, Matthews correlation coefficient

MDA, mixture discriminant analysis

ML, machine learning

MLP, multi-layer perceptron

MRI, magnetic resonance imaging

NB, naive Bayes

NGLDM, neighbouring gray-level dependence matrix

NGTDM, neighbouring gray-tone difference matrix

NNET, neural network

NZV, near zero variance

OPC, oropharyngeal cancer

OPSCC, oropharyngeal squamous cell carcinoma

OS, overall survival

OSCC, oral squamous cell carcinoma

PCA, principal component analysis

PET, positron emission tomography

PG-SGA, patient-generated subjective global assessment

PLS, partial least squares

QDA, quadratic discriminant analysis

QoL, quality of life

RF, random forest

ROC curve, receiver operating characteristic curve

ROI, region of interest

ROSE, random over sampling examples

SE, standard error

SMFD, skeletal muscle function deficit

SMI, skeletal mass index

SMOTE, synthetic minority oversampling technique

SMR, skeletal muscle radiodensity

SR, super resolution

SVM, support vector machine

T2, thoracic vertebra 2

t-SNE, t-distributed stochastic neighbor embedding

UMAP, uniform manifold approximation and projection

WHO PS, world health organization performance status

ZV, zero variance

REFERENCES

[1] J. D. Cramer, B. Burtness, Q. T. Le, and R. L. Ferris, "The changing therapeutic landscape of head and neck cancer," *Nat Rev Clin Oncol*, vol. 16, no. 11, pp. 669–683, Nov. 2019, doi: 10.1038/s41571-019-0227-z.

[2] S. Cavalieri *et al.*, "Development of a multiomics database for personalized prognostic forecasting in head and neck cancer: The Big Data to Decide <scp>EU</scp> Project," *Head Neck*, vol. 43, no. 2, pp. 601–612, Feb. 2021, doi: 10.1002/hed.26515.

[3] D. E. Johnson, B. Burtness, C. R. Leemans, V. W. Y. Lui, J. E. Bauman, and J. R. Grandis, "Head and neck squamous cell carcinoma," *Nat Rev Dis Primers*, vol. 6, no. 1, p. 92, Dec. 2020, doi: 10.1038/s41572-020-00224-3.

[4] Y.-P. Chen, A. T. C. Chan, Q.-T. Le, P. Blanchard, Y. Sun, and J. Ma, "Nasopharyngeal carcinoma.," *Lancet*, vol. 394, no. 10192, pp. 64–80, Jul. 2019, doi: 10.1016/S0140-6736(19)30956-0.

[5] J. S. Lewis, "Sinonasal Squamous Cell Carcinoma: A Review with Emphasis on Emerging Histologic Subtypes and the Role of Human Papillomavirus," *Head Neck Pathol*, vol. 10, no. 1, pp. 60–67, Mar. 2016, doi: 10.1007/s12105-016-0692-y.

[6] E. P. Simard, L. A. Torre, and A. Jemal, "International trends in head and neck cancer incidence rates: differences by country, sex and anatomic site.," *Oral Oncol*, vol. 50, no. 5, pp. 387–403, May 2014, doi: 10.1016/j.oraloncology.2014.01.016.

[7] S. Marur and A. A. Forastiere, "Head and Neck Squamous Cell Carcinoma: Update on Epidemiology, Diagnosis, and Treatment," *Mayo Clin Proc*, vol. 91, no. 3, pp. 386–396, Mar. 2016, doi: 10.1016/j.mayocp.2015.12.017.

[8] M. L. Gillison, A. K. Chaturvedi, W. F. Anderson, and C. Fakhry, "Epidemiology of Human Papillomavirus–Positive Head and Neck Squamous Cell Carcinoma," *Journal of Clinical Oncology*, vol. 33, no. 29, pp. 3235–3242, Oct. 2015, doi: 10.1200/JCO.2015.61.6995.

[9] H. Mehanna *et al.*, "Geographic variation in human papillomavirus-related oropharyngeal cancer: Data from 4 multinational randomized trials," *Head Neck*, vol. 38, no. S1, pp. E1863–E1869, Apr. 2016, doi: 10.1002/hed.24336.

[10] E. M. Rettig and G. D'Souza, "Epidemiology of head and neck cancer.," *Surg Oncol Clin N Am*, vol. 24, no. 3, pp. 379–96, Jul. 2015, doi: 10.1016/j.soc.2015.03.001.

[11] J. D. McDermott and D. W. Bowles, "Epidemiology of Head and Neck Squamous Cell Carcinomas: Impact on Staging and Prevention Strategies," *Curr Treat Options Oncol*, vol. 20, no. 5, p. 43, May 2019, doi: 10.1007/s11864-019-0650-5.

[12] N. McQueen, E. J. Partington, K. F. Harrington, E. L. Rosenthal, W. R. Carroll, and C. E. Schmalbach, "Smoking Cessation and Electronic Cigarette Use among Head and Neck Cancer Patients," *Otolaryngology–Head and Neck Surgery*, vol. 154, no. 1, pp. 73–79, Jan. 2016, doi: 10.1177/0194599815613279.

[13] V. Yu *et al.*, "Electronic cigarettes induce DNA strand breaks and cell death independently of nicotine in cell lines.," *Oral Oncol*, vol. 52, pp. 58–65, Jan. 2016, doi: 10.1016/j.oraloncology.2015.10.018.

[14] B. A. Primack *et al.*, "Initiation of Traditional Cigarette Smoking after Electronic Cigarette Use Among Tobacco-Naïve US Young Adults," *Am J Med*, vol. 131, no. 4, pp. 443.e1-443.e9, Apr. 2018, doi: 10.1016/j.amjmed.2017.11.005.

[15] A. J. Cruz-Jentoft *et al.*, "Sarcopenia: revised European consensus on definition and diagnosis," *Age Ageing*, vol. 48, no. 1, pp. 16–31, Jan. 2019, doi: 10.1093/ageing/afy169.

[16] R. M. Dodds *et al.*, "Grip Strength across the Life Course: Normative Data from Twelve British Studies," *PLoS One*, vol. 9, no. 12, p. e113637, Dec. 2014, doi: 10.1371/journal.pone.0113637.

[17] M. Cesari *et al.*, "Added Value of Physical Performance Measures in Predicting Adverse Health-Related Events: Results from the Health, Aging and Body Composition Study," *J Am Geriatr Soc*, vol. 57, no. 2, pp. 251–259, Feb. 2009, doi: 10.1111/j.1532-5415.2008.02126.x.

[18] S. A. Studenski *et al.*, "The FNIH sarcopenia project: rationale, study description, conference recommendations, and final estimates.," *J Gerontol A Biol Sci Med Sci*, vol. 69, no. 5, pp. 547–58, May 2014, doi: 10.1093/gerona/glu010.

[19] H. Gould, S. L. Brennan, M. A. Kotowicz, G. C. Nicholson, and J. A. Pasco, "Total and appendicular lean mass reference ranges for Australian men and women: the Geelong osteoporosis study.," *Calcif Tissue Int*, vol. 94, no. 4, pp. 363–72, Apr. 2014, doi: 10.1007/s00223-013-9830-7.

[20] A. J. Cruz-Jentoft *et al.*, "Sarcopenia: European consensus on definition and diagnosis: Report of the European Working Group on Sarcopenia in Older People.," *Age Ageing*, vol. 39, no. 4, pp. 412–23, Jul. 2010, doi: 10.1093/ageing/afq034.

[21] S. Studenski *et al.*, "Gait speed and survival in older adults.," *JAMA*, vol. 305, no. 1, pp. 50–8, Jan. 2011, doi: 10.1001/jama.2010.1923.

[22] R. Pavasini *et al.*, "Short Physical Performance Battery and all-cause mortality: systematic review and meta-analysis," *BMC Med*, vol. 14, no. 1, p. 215, Dec. 2016, doi: 10.1186/s12916-016-0763-7.

[23] J. M. Guralnik, L. Ferrucci, E. M. Simonsick, M. E. Salive, and R. B. Wallace, "Lower-extremity function in persons over the age of 70 years as a predictor of subsequent disability.," *N Engl J Med*, vol. 332, no. 9, pp. 556–61, Mar. 1995, doi: 10.1056/NEJM199503023320902.

[24] H. A. Bischoff *et al.*, "Identifying a cut-off point for normal mobility: a comparison of the timed 'up and go' test in community-dwelling and institutionalised elderly women.," *Age Ageing*, vol. 32, no. 3, pp. 315–20, May 2003, doi: 10.1093/ageing/32.3.315.

[25] A. B. Newman *et al.*, "Association of Long-Distance Corridor Walk Performance With Mortality, Cardiovascular Disease, Mobility Limitation, and Disability," *JAMA*, vol. 295, no. 17, p. 2018, May 2006, doi: 10.1001/jama.295.17.2018.

[26] A. Clegg, J. Young, S. Iliffe, M. O. Rikkert, and K. Rockwood, "Frailty in elderly people.," *Lancet*, vol. 381, no. 9868, pp. 752–62, Mar. 2013, doi: 10.1016/S0140-6736(12)62167-9.

[27] J. E. Morley *et al.*, "Frailty consensus: a call to action.," *J Am Med Dir Assoc*, vol. 14, no. 6, pp. 392–7, Jun. 2013, doi: 10.1016/j.jamda.2013.03.022.

[28] T. S. Fu *et al.*, "Is Frailty Associated With Worse Outcomes After Head and Neck Surgery? A Narrative Review," *Laryngoscope*, vol. 130, no. 6, pp. 1436–1442, Jun. 2020, doi: 10.1002/lary.28307.

[29] L. P. Fried *et al.*, "Frailty in older adults: evidence for a phenotype.," *J Gerontol A Biol Sci Med Sci*, vol. 56, no. 3, pp. M146-56, Mar. 2001, doi: 10.1093/gerona/56.3.m146.

[30] K. Rockwood and A. Mitnitski, "Frailty in relation to the accumulation of deficits.," *J Gerontol A Biol Sci Med Sci*, vol. 62, no. 7, pp. 722–7, Jul. 2007, doi: 10.1093/gerona/62.7.722.

[31] S. G. Parker *et al.*, "What is Comprehensive Geriatric Assessment (CGA)? An umbrella review.," *Age Ageing*, vol. 47, no. 1, pp. 149–155, 2018, doi: 10.1093/ageing/afx166.

[32] M. E. Hamaker, J. M. Jonker, S. E. de Rooij, A. G. Vos, C. H. Smorenburg, and B. C. van Munster, "Frailty screening methods for predicting outcome of a comprehensive geriatric assessment in elderly patients with cancer: a systematic review," *Lancet Oncol*, vol. 13, no. 10, pp. e437–e444, Oct. 2012, doi: 10.1016/S1470-2045(12)70259-0.

[33] P. Soubeyran *et al.*, "Validation of a screening test for elderly patients in oncology," *Journal of Clinical Oncology*, vol. 26, no. 15_suppl, pp. 20568–20568, May 2008, doi: 10.1200/jco.2008.26.15_suppl.20568.

[34] R. Correa-de-Araujo *et al.*, "Myosteatosis in the Context of Skeletal Muscle Function Deficit: An Interdisciplinary Workshop at the National Institute on Aging.," *Front Physiol*, vol. 11, p. 963, 2020, doi: 10.3389/fphys.2020.00963.

[35] R. Correa-de-Araujo, M. O. Harris-Love, I. Miljkovic, M. S. Fragala, B. W. Anthony, and T. M. Manini, "The Need for Standardized Assessment of Muscle Quality in Skeletal Muscle Function Deficit and Other Aging-Related Muscle Dysfunctions: A Symposium Report.," *Front Physiol*, vol. 8, p. 87, 2017, doi: 10.3389/fphys.2017.00087.

[36] J. Aubrey *et al.*, "Measurement of skeletal muscle radiation attenuation and basis of its biological variation.," *Acta Physiol (Oxf)*, vol. 210, no. 3, pp. 489–97, Mar. 2014, doi: 10.1111/apha.12224.

[37] R. J. Gillies, P. E. Kinahan, and H. Hricak, "Radiomics: Images Are More than Pictures, They Are Data.," *Radiology*, vol. 278, no. 2, pp. 563–77, Feb. 2016, doi: 10.1148/radiol.2015151169.

[38] J. E. van Timmeren, D. Cester, S. Tanadini-Lang, H. Alkadhi, and B. Baessler, "Radiomics in medical imaging—'how-to' guide and critical reflection," *Insights Imaging*, vol. 11, no. 1, p. 91, Dec. 2020, doi: 10.1186/s13244-020-00887-2.

[39] J.-E. Bibault *et al.*, "Radiomics: A primer for the radiation oncologist.," *Cancer Radiother*, vol. 24, no. 5, pp. 403–410, Aug. 2020, doi: 10.1016/j.canrad.2020.01.011.

[40] G. Bruixola *et al.*, "Radiomics and radiogenomics in head and neck squamous cell carcinoma: Potential contribution to patient management and challenges.," *Cancer Treat Rev*, vol. 99, p. 102263, Sep. 2021, doi: 10.1016/j.ctrv.2021.102263.

[41] M. Piñeiro-Fiel, A. Moscoso, V. Pubul, Á. Ruibal, J. Silva-Rodríguez, and P. Aguiar, "A Systematic Review of PET Textural Analysis and Radiomics in Cancer.," *Diagnostics (Basel)*, vol. 11, no. 2, Feb. 2021, doi: 10.3390/diagnostics11020380.

[42] A. Zwanenburg *et al.*, "The Image Biomarker Standardization Initiative: Standardized Quantitative Radiomics for High-Throughput Image-based Phenotyping," *Radiology*, vol. 295, no. 2, pp. 328–338, May 2020, doi: 10.1148/radiol.2020191145.

[43] L. G. Marcu, C. Boyd, and E. Bezak, "Feeding the Data Monster: Data Science in Head and Neck Cancer for Personalized Therapy.," *J Am Coll Radiol*, vol. 16, no. 12, pp. 1695–1701, Dec. 2019, doi: 10.1016/j.jacr.2019.05.045.

[44] M. Findlay, K. White, N. Stapleton, and J. Bauer, "Is sarcopenia a predictor of prognosis for patients undergoing radiotherapy for head and neck cancer? A meta-analysis," *Clinical Nutrition*, vol. 40, no. 4, pp. 1711–1718, Apr. 2021, doi: 10.1016/j.clnu.2020.09.017.

[45] A. Surov and A. Wienke, "Low skeletal muscle mass predicts relevant clinical outcomes in head and neck squamous cell carcinoma. A meta analysis," *Ther Adv Med Oncol*, vol. 13, p. 175883592110088, Jan. 2021, doi: 10.1177/17588359211008844.

[46] M. Findlay, K. White, M. Lai, D. Luo, and J. D. Bauer, "The Association Between Computed Tomography–Defined Sarcopenia and Outcomes in Adult Patients Undergoing Radiotherapy of Curative Intent for Head and Neck Cancer: A Systematic Review," *J Acad Nutr Diet*, vol. 120, no. 8, pp. 1330-1347.e8, Aug. 2020, doi: 10.1016/j.jand.2020.03.021.

[47] M. I. van Rijn-Dekker *et al.*, "Impact of sarcopenia on survival and late toxicity in head and neck cancer patients treated with radiotherapy," *Radiotherapy and Oncology*, vol. 147, pp. 103–110, Jun. 2020, doi: 10.1016/j.radonc.2020.03.014.

[48] A. Tamaki, N. F. Manzoor, E. Babajanian, M. Ascha, R. Rezaee, and C. A. Zender, "Clinical Significance of Sarcopenia among Patients with Advanced Oropharyngeal Cancer," *Otolaryngology–Head and Neck Surgery*, vol. 160, no. 3, pp. 480–487, Mar. 2019, doi: 10.1177/0194599818793857.

[49] M. Findlay, C. Brown, R. De Abreu Lourenço, K. White, and J. Bauer, "Sarcopenia and myosteatosis in patients undergoing curative radiotherapy for head and neck cancer: Impact on survival, treatment completion, hospital admission and cost," *Journal of Human Nutrition and Dietetics*, vol. 33, no. 6, pp. 811–821, Dec. 2020, doi: 10.1111/jhn.12788.

[50] L. Martin *et al.*, "Cancer Cachexia in the Age of Obesity: Skeletal Muscle Depletion Is a Powerful Prognostic Factor, Independent of Body Mass Index," *Journal of Clinical Oncology*, vol. 31, no. 12, pp. 1539–1547, Apr. 2013, doi: 10.1200/JCO.2012.45.2722.

[51] E. Ahern *et al.*, "Impact of sarcopenia and myosteatosis on survival outcomes for patients with head and neck cancer undergoing curative-intent treatment," *British Journal of Nutrition*, vol. 129, no. 3, pp. 406–415, Feb. 2023, doi: 10.1017/S0007114522000435.

[52] F. J. van Deudekom, A. S. Schimberg, M. H. Kallenberg, M. Slingerland, L.-A. van der Velden, and S. P. Mooijaart, "Functional and cognitive impairment, social environment, frailty and adverse health outcomes in older patients with head and neck cancer, a systematic review.," *Oral Oncol*, vol. 64, pp. 27–36, 2017, doi: 10.1016/j.oraloncology.2016.11.013.

[53] D. P. Goldstein *et al.*, "Frailty as a predictor of outcomes in patients undergoing head and neck cancer surgery," *Laryngoscope*, vol. 130, no. 5, May 2020, doi: 10.1002/lary.28222.

[54] M. S. Crestani, T. Grassi, and T. Steemburgo, "Methods of nutritional assessment and functional capacity in the identification of unfavorable clinical outcomes in hospitalized patients with cancer: a systematic review," *Nutr Rev*, vol. 80, no. 4, pp. 786–811, Mar. 2022, doi: 10.1093/nutrit/nuab090.

[55] M. Findlay, K. White, C. Brown, and J. D. Bauer, "Nutritional status and skeletal muscle status in patients with head and neck cancer: Impact on outcomes," *J Cachexia Sarcopenia Muscle*, vol. 12, no. 6, pp. 2187–2198, Dec. 2021, doi: 10.1002/jcsm.12829.

[56] D. Albano, C. Messina, J. Vitale, and L. M. Sconfienza, "Imaging of sarcopenia: old evidence and new insights," *Eur Radiol*, vol. 30, no. 4, pp. 2199–2208, Apr. 2020, doi: 10.1007/s00330-019-06573-2.

[57] W. Shen *et al.*, "Total body skeletal muscle and adipose tissue volumes: estimation from a single abdominal cross-sectional image," *J Appl Physiol*, vol. 97, no. 6, pp. 2333–2338, Dec. 2004, doi: 10.1152/japplphysiol.00744.2004.

[58] J. E. Swartz *et al.*, "Feasibility of using head and neck CT imaging to assess skeletal muscle mass in head and neck cancer patients," *Oral Oncol*, vol. 62, pp. 28–33, Nov. 2016, doi: 10.1016/j.oraloncology.2016.09.006.

[59] M. J. Page *et al.*, "The PRISMA 2020 statement: an updated guideline for reporting systematic reviews," *BMJ*, p. n71, Mar. 2021, doi: 10.1136/bmj.n71.

[60] N. Fiorini, D. J. Lipman, and Z. Lu, "Towards PubMed 2.0," *Elife*, vol. 6, Oct. 2017, doi: 10.7554/eLife.28801.

[61] F. Ufuk, D. Herek, and D. Yüksel, "Diagnosis of Sarcopenia in Head and Neck Computed Tomography: Cervical Muscle Mass as a Strong Indicator of Sarcopenia," *Clin Exp Otorhinolaryngol*, vol. 12, no. 3, pp. 317–324, Aug. 2019, doi: 10.21053/ceo.2018.01613.

[62] B. Vangelov, J. Bauer, D. Moses, and R. Smee, "A prediction model for skeletal muscle evaluation and computed tomography-defined sarcopenia diagnosis in a predominantly overweight cohort of patients with head and neck cancer," *European Archives of Oto-Rhino-Laryngology*, vol. 280, no. 1, pp. 321–328, Jan. 2023, doi: 10.1007/s00405-022-07545-x.

[63] J. E. Swartz *et al.*, "Feasibility of using head and neck CT imaging to assess skeletal muscle mass in head and neck cancer patients," *Oral Oncol*, vol. 62, pp. 28–33, Nov. 2016, doi: 10.1016/j.oraloncology.2016.09.006.

[64] X. Lu *et al.*, "Evaluating the prognosis of oral squamous cell carcinoma patients via L3 skeletal muscle index," *Oral Dis*, vol. 29, no. 3, pp. 923–932, Apr. 2023, doi: 10.1111/odi.14074.

[65] B. Olson *et al.*, "Establishment and Validation of Pre-Therapy Cervical Vertebrae Muscle Quantification as a Prognostic Marker of Sarcopenia in Patients With Head and Neck Cancer," *Front Oncol*, vol. 12, Feb. 2022, doi: 10.3389/fonc.2022.812159.

[66] B. Vangelov, J. Bauer, D. Moses, and R. Smee, "The effectiveness of skeletal muscle evaluation at the third cervical vertebral level for computed tomography-defined sarcopenia assessment in patients with head and neck cancer," *Head Neck*, vol. 44, no. 5, pp. 1047–1056, May 2022, doi: 10.1002/hed.27000.

[67] S. I. Bril *et al.*, "Validation of skeletal muscle mass assessment at the level of the third cervical vertebra in patients with head and neck cancer," *Oral Oncol*, vol. 123, p. 105617, Dec. 2021, doi: 10.1016/j.oraloncology.2021.105617.

[68] J.-K. Yoon, J. Y. Jang, Y.-S. An, and S. J. Lee, "Skeletal muscle mass at C3 may not be a strong predictor for skeletal muscle mass at L3 in sarcopenic patients with head and neck cancer," *PLoS One*, vol. 16, no. 7, p. e0254844, Jul. 2021, doi: 10.1371/journal.pone.0254844.

[69] A. R. Jung, J.-L. Roh, J. S. Kim, S.-H. Choi, S. Y. Nam, and S. Y. Kim, "Efficacy of head and neck computed tomography for skeletal muscle mass estimation in patients with head and neck cancer," *Oral Oncol*, vol. 95, pp. 95–99, Aug. 2019, doi: 10.1016/j.oraloncology.2019.06.009.

[70] B. Vangelov, J. Bauer, D. Moses, and R. Smee, "The use of the second thoracic vertebral landmark for skeletal muscle assessment and computed tomography-defined sarcopenia evaluation in patients with head and neck cancer," *Head Neck*, vol. 45, no. 4, pp. 1006–1016, Apr. 2023, doi: 10.1002/hed.27320.

[71] H. C. van Heusden *et al.*, "Feasibility of assessment of skeletal muscle mass on a single cross-sectional image at the level of the fourth thoracic vertebra," *Eur J Radiol*, vol. 142, p. 109879, Sep. 2021, doi: 10.1016/j.ejrad.2021.109879.

[72] R. Matsuyama *et al.*, "Assessing skeletal muscle mass based on the cross-sectional area of muscles at the 12th thoracic vertebra level on computed tomography in patients with oral squamous cell carcinoma," *Oral Oncol*, vol. 113, p. 105126, Feb. 2021, doi: 10.1016/j.oraloncology.2020.105126.

[73] H. C. van Heusden, N. Chargi, J. W. Dankbaar, E. J. Smid, and R. de Bree, "Masseter muscle parameters can function as an alternative for skeletal muscle mass assessments on cross-sectional imaging at lumbar or cervical vertebral levels," *Quant Imaging Med Surg*, vol. 12, no. 1, pp. 15–27, Jan. 2022, doi: 10.21037/qims-21-43.

[74] S.-W. Chang *et al.*, "Masticatory muscle index for indicating skeletal muscle mass in patients with head and neck cancer," *PLoS One*, vol. 16, no. 5, p. e0251455, May 2021, doi: 10.1371/journal.pone.0251455.

[75] D. Yunaiyama *et al.*, "Sarcopenia at the infrahyoid level as a prognostic factor in patients with advanced-stage non-virus-related head and neck carcinoma," *European Archives of Oto-Rhino-Laryngology*, vol. 279, no. 6, pp. 3131–3137, Jun. 2022, doi: 10.1007/s00405-021-07147-z.

[76] R. Bentahila *et al.*, "The impact of sarcopenia on survival and treatment tolerance in patients with head and neck cancer treated with chemoradiotherapy," *Cancer Med*, vol. 12, no. 4, pp. 4170–4183, Feb. 2023, doi: 10.1002/cam4.5278.

[77] Ah. R. Jung, J.-L. Roh, J. S. Kim, S.-H. Choi, S. Y. Nam, and S. Y. Kim, "The impact of skeletal muscle depletion on older adult patients with head and neck cancer undergoing primary surgery," *J Geriatr Oncol*, vol. 12, no. 1, pp. 128–133, Jan. 2021, doi: 10.1016/j.jgo.2020.06.009.

[78] C. M. Prado *et al.*, "Prevalence and clinical implications of sarcopenic obesity in patients with solid tumours of the respiratory and gastrointestinal tracts: a population-based study," *Lancet Oncol*, vol. 9, no. 7, pp. 629–635, Jul. 2008, doi: 10.1016/S1470-2045(08)70153-0.

[79] M. Fattouh *et al.*, "Association between pretreatment obesity, sarcopenia, and survival in patients with head and neck cancer," *Head Neck*, vol. 41, no. 3, pp. 707–714, Mar. 2019, doi: 10.1002/hed.25420.

[80] A. R. Jung *et al.*, "Prognostic value of body composition on recurrence and survival of advanced-stage head and neck cancer," *Eur J Cancer*, vol. 116, pp. 98–106, Jul. 2019, doi: 10.1016/j.ejca.2019.05.006.

[81] Y. Cho, J. W. Kim, K. C. Keum, C. G. Lee, H. C. Jeung, and I. J. Lee, "Prognostic Significance of Sarcopenia With Inflammation in Patients With Head and Neck Cancer Who Underwent Definitive Chemoradiotherapy," *Front Oncol*, vol. 8, Oct. 2018, doi: 10.3389/fonc.2018.00457.

[82] Y.-S. Kim *et al.*, "Prevalence of Sarcopenia and Sarcopenic Obesity in the Korean Population Based on the Fourth Korean National Health and Nutritional Examination Surveys," *J Gerontol A Biol Sci Med Sci*, vol. 67, no. 10, pp. 1107–1113, Oct. 2012, doi: 10.1093/gerona/gls071.

[83] S.-I. Go *et al.*, "Sarcopenia and inflammation are independent predictors of survival in male patients newly diagnosed with small cell lung cancer," *Supportive Care in Cancer*, vol. 24, no. 5, pp. 2075–2084, May 2016, doi: 10.1007/s00520-015-2997-x.

[84] R. Kabarriti *et al.*, "The impact of dietary regimen compliance on outcomes for HNSCC patients treated with radiation therapy," *Supportive Care in Cancer*, vol. 26, no. 9, pp. 3307–3313, Sep. 2018, doi: 10.1007/s00520-018-4198-x.

[85] A. J. Grossberg *et al.*, "Association of Body Composition With Survival and Locoregional Control of Radiotherapy-Treated Head and Neck Squamous Cell Carcinoma," *JAMA Oncol*, vol. 2, no. 6, p. 782, Jun. 2016, doi: 10.1001/jamaoncol.2015.6339.

[86] P. Bonavolontà *et al.*, "Evaluation of sarcopenia and sarcopenic obesity in patients affected by oral squamous cell carcinoma: A retrospective single-center study," *Journal of Cranio-Maxillofacial Surgery*, vol. 51, no. 1, pp. 7–15, Jan. 2023, doi: 10.1016/j.jcms.2023.01.014.

[87] A. W. Wendrich *et al.*, "Low skeletal muscle mass is a predictive factor for chemotherapy dose-limiting toxicity in patients with locally advanced head and neck cancer," *Oral Oncol*, vol. 71, pp. 26–33, Aug. 2017, doi: 10.1016/j.oraloncology.2017.05.012.

[88] P. Nagpal, D. S. Pruthi, M. Pandey, A. Yadav, and H. Singh, "Impact of sarcopenia in locally advanced head and neck cancer treated with chemoradiation: An Indian tertiary care hospital experience," *Oral Oncol*, vol. 121, p. 105483, Oct. 2021, doi: 10.1016/j.oraloncology.2021.105483.

[89] N. Chargi, S. I. Bril, M. H. Emmelot-Vonk, and R. de Bree, "Sarcopenia is a prognostic factor for overall survival in elderly patients with head-and-neck cancer," *European Archives of Oto-Rhino-Laryngology*, vol. 276, no. 5, pp. 1475–1486, May 2019, doi: 10.1007/s00405-019-05361-4.

[90] K. Yamahara, A. Mizukoshi, K. Lee, and S. Ikegami, "Sarcopenia with inflammation as a predictor of survival in patients with head and neck cancer," *Auris Nasus Larynx*, vol. 48, no. 5, pp. 1013–1022, Oct. 2021, doi: 10.1016/j.anl.2021.03.021.

[91] J. Lee *et al.*, "Sarcopenia and Systemic Inflammation Synergistically Impact Survival in Oral Cavity Cancer," *Laryngoscope*, vol. 131, no. 5, May 2021, doi: 10.1002/lary.29221.

[92] R. G. Ganju, R. Morse, A. Hoover, M. TenNapel, and C. E. Lominska, "The impact of sarcopenia on tolerance of radiation and outcome in patients with head and neck cancer receiving chemoradiation," *Radiotherapy and Oncology*, vol. 137, pp. 117–124, Aug. 2019, doi: 10.1016/j.radonc.2019.04.023.

[93] D. M. McGoldrick, A. Yassin Alsabbagh, M. Shaikh, L. Pettit, and S. K. Bhatia, "Masseter muscle defined sarcopenia and survival in head and neck cancer patients," *British Journal of Oral and Maxillofacial Surgery*, vol. 60, no. 4, pp. 454–458, May 2022, doi: 10.1016/j.bjoms.2021.07.020.

[94] T. Chinnery *et al.*, "Utilizing Artificial Intelligence for Head and Neck Cancer Outcomes Prediction From Imaging," *Canadian Association of Radiologists Journal*, vol. 72, no. 1, pp. 73–85, Feb. 2021, doi: 10.1177/0846537120942134.

[95] S. Volpe *et al.*, "Machine Learning for Head and Neck Cancer: A Safe Bet?—A Clinically Oriented Systematic Review for the Radiation Oncologist," *Front Oncol*, vol. 11, Nov. 2021, doi: 10.3389/fonc.2021.772663.

[96] L. V. van Dijk and C. D. Fuller, "Artificial Intelligence and Radiomics in Head and Neck Cancer Care: Opportunities, Mechanics, and Challenges," *American Society of Clinical Oncology Educational Book*, no. 41, pp. e225–e235, Jun. 2021, doi: 10.1200/EDBK_320951.

[97] W. Luo *et al.*, "Guidelines for Developing and Reporting Machine Learning Predictive Models in Biomedical Research: A Multidisciplinary View," *J Med Internet Res*, vol. 18, no. 12, p. e323, Dec. 2016, doi: 10.2196/jmir.5870.

[98] K. Clark *et al.*, "The Cancer Imaging Archive (TCIA): Maintaining and Operating a Public Information Repository," *J Digit Imaging*, vol. 26, no. 6, pp. 1045–1057, Dec. 2013, doi: 10.1007/s10278-013-9622-7.

[99] M. Vallières *et al.*, "Radiomics strategies for risk assessment of tumour failure in head-and-neck cancer," *Sci Rep*, vol. 7, no. 1, p. 10117, Aug. 2017, doi: 10.1038/s41598-017-10371-5.

[100] R Core Team, "R: A Language and Environment for Statistical Computing." Vienna, Austria, 2021. [Online]. Available: https://www.R-project.org/

[101] T. M. Therneau, "survival: Survival Analysis." 2021. [Online]. Available: https://github.com/therneau/survival

[102] A. Kassambara, M. Kosinski, and P. Biecek, "survminer: Drawing Survival Curves using ggplot2." 2021. [Online]. Available: https://rpkgs.datanovia.com/survminer/index.html

[103] D. Datta, "blandr: a Bland-Altman Method Comparison package for R." 2017. doi: 10.5281/zenodo.824514.

[104] W. Revelle, "psych: Procedures for Psychological, Psychometric, and Personality Research." 2021. [Online]. Available: https://personality-project.org/r/psych/

[105] B. Hamner and M. Frasco, "Metrics: Evaluation Metrics for Machine Learning." 2018. [Online]. Available: https://github.com/mfrasco/Metrics

[106] A. Fedorov *et al.*, "3D Slicer as an image computing platform for the Quantitative Imaging Network.," *Magn Reson Imaging*, vol. 30, no. 9, pp. 1323–41, Nov. 2012, doi: 10.1016/j.mri.2012.05.001.

[107] J. J. M. van Griethuysen *et al.*, "Computational Radiomics System to Decode the Radiographic Phenotype," *Cancer Res*, vol. 77, no. 21, pp. e104–e107, Nov. 2017, doi: 10.1158/0008-5472.CAN-17-0339.

[108] Bradski G., "The OpenCV Library," *Dr. Dobb's Journal of Software Tools*, vol. 120, pp. 122–125, 2000, [Online]. Available: http://www.drdobbs.com/open-source/the-opencv-library/184404319

[109] S. van der Walt *et al.*, "scikit-image: image processing in Python," *PeerJ*, vol. 2, p. e453, Jun. 2014, doi: 10.7717/peerj.453.

[110] J. B. Ewaschuk, A. Almasud, and V. C. Mazurak, "Role of n-3 fatty acids in muscle loss and myosteatosis," *Applied Physiology, Nutrition, and Metabolism*, vol. 39, no. 6, pp. 654–662, Jun. 2014, doi: 10.1139/apnm-2013-0423.

[111] P. Mildenberger, M. Eichelberg, and E. Martin, "Introduction to the DICOM standard," *Eur Radiol*, vol. 12, no. 4, pp. 920–927, Apr. 2002, doi: 10.1007/s003300101100.

[112] T. K. Koo and M. Y. Li, "A Guideline of Selecting and Reporting Intraclass Correlation Coefficients for Reliability Research," *J Chiropr Med*, vol. 15, no. 2, pp. 155–163, Jun. 2016, doi: 10.1016/j.jcm.2016.02.012.

[113] Y. Zhang, A. Oikonomou, A. Wong, M. A. Haider, and F. Khalvati, "Radiomics-based Prognosis Analysis for Non-Small Cell Lung Cancer," *Sci Rep*, vol. 7, no. 1, p. 46349, Apr. 2017, doi: 10.1038/srep46349.

[114] H. J. W. L. Aerts *et al.*, "Decoding tumour phenotype by noninvasive imaging using a quantitative radiomics approach," *Nat Commun*, vol. 5, no. 1, p. 4006, Jun. 2014, doi: 10.1038/ncomms5006.

[115] R. Reiazi *et al.*, "The impact of the variation of imaging parameters on the robustness of Computed Tomography radiomic features: A review," *Comput Biol Med*, vol. 133, p. 104400, Jun. 2021, doi: 10.1016/j.compbiomed.2021.104400.

[116] X. Teng *et al.*, "Building reliable radiomic models using image perturbation," *Sci Rep*, vol. 12, no. 1, p. 10035, Jun. 2022, doi: 10.1038/s41598-022-14178-x.

[117] X. Teng *et al.*, "Improving radiomic model reliability using robust features from perturbations for head-and-neck carcinoma," *Front Oncol*, vol. 12, Oct. 2022, doi: 10.3389/fonc.2022.974467.

[118] E. C. de Farias, C. di Noia, C. Han, E. Sala, M. Castelli, and L. Rundo, "Impact of GAN-based lesion-focused medical image super-resolution on the robustness of radiomic features," *Sci Rep*, vol. 11, no. 1, p. 21361, Nov. 2021, doi: 10.1038/s41598-021-00898-z.

[119] van der Maaten L.J.P. and G. E. Hinton, "Visualizing High-Dimensional Data Using t-SNE," *Journal of Machine Learning Research*, vol. 9, no. Nov, pp. 2579–2605, 2008.

[120] L. McInnes, J. Healy, and J. Melville, "UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction," Feb. 2018.

[121] F. Pedregosa *et al.*, "Scikit-learn: Machine learning in Python," *Journal of Machine Learning Research*, vol. 12, no. Oct, pp. 2825–2830, 2011, [Online]. Available: https://scikit-learn.org/stable/about.html

[122] Mavs, "ATOM: A Python package for fast exploration of machine learning pipelines." Nov. 2019. [Online]. Available: https://tvdboom.github.io/ATOM/

[123] C. K. A. Guillaume Lemaitre Fernando Nogueira, "Imbalanced-learn: A Python Toolbox to Tackle the Curse of Imbalanced Datasets in Machine Learning," *Journal of Machine Learning Research*, vol. 18, no. 17, pp. 1–5, 2017, [Online]. Available: http://jmlr.org/papers/v18/16-365.html