# Comprehensive, Lesion-level, Longitudinal Assessment of Metastatic Cancer, and its Clinical Implications

By

Victor Santoro-Fernandes

A dissertation submitted in partial fulfillment of
the requirements for the degree of

Doctor of Philosophy
(Medical Physics)

at the
UNIVERSITY OF WISCONSIN-MADISON
2023

Date of final oral examination: 11/17/2023
The dissertation is subject to approval by the following members of the Final Oral Committe:

Robert Jeraj, Professor, Medical Physics
Oliver Wieben, Professor, Medical Physics
Diego Hernando, Associate Professor, Medical Physics and Radiology
Glenn Liu, Professor, Medicine (Hematology/Oncology) and Medical Physics
John Floberg, Assistant Professor, Medicine (Human Oncology)

# ACKNOWLEDGEMENTS

I thank Dr. Robert Jeraj for his wise and unconventionally kind mentorship that never ceases to surprise me, impress me, and inspire me to be the best person I can possibly be. I also thank Dr. Paulo Costa for introducing me to the field of Medical Physics and to Dr. Robert Jeraj.

This dissertation's work is inserted in the ethos of our research group, which I treasured so much during my years spent here. For that I thank my colleagues and friends Dr. Alison Deatsch, Dr. Amy Weisman, Dr. Peter Ferjančič, Dr. Daniel Huff, Dr. Alison Roth, Dr. Tim Perk, Brayden Schott, Dr. Stephen Yip, Dr. Christie Lin, Dr. Mathew Scarpelli, Dr. Mauro Namías, Žan Klaneček, Katja Strašek, and Luciano Rivetti. I am also thankful for the help of our unexpected visitors, Thomas Francken, Ryan Meeker, Angela Galindo, and Louis Nielsen. My gratitude also goes to our colleagues and group members from across the pond, for their valuable contributions through so many exciting group meetings, namely Dr. Martin Ebert, Dr. Andrej Studen, Dr. Urban Simončič, Dr. Matija Milanič, Dr. Martina Reberšek, Dr. Nežka Hribernik, Dr. Damijan Valentinuzzi, Dr. Maruša Turk, Jošt Stergar, Dr. Matej Perovnik, Matic Orehar, Tadej Tomanič, Roc Dolenec, Živa Drakulić, Dr. Gašper Razdevšek, Dr. Marija Delić, Sergei Diuzhenko, Eva Rebec, and Doruntina Hoxha.

I thank Dr. Glenn Liu, Dr. Diego Hernando, Dr. John Floberg, Dr. Oliver Wieben, Dr. Scott Perlman, and Dr. Sean Fain, for being part of my Ph.D. committee. For their clinical expertise and input, I thank Dr. Quinton Keigley, Dr. Mark Albertini, Dr. Vince Ma, Dr. Nathalya Uboha, Dr. Steve Cho, Dr. Renuka Iyer, and Dr. Christos Fountzilas, the clinicians with whom I had the pleasure to collaborate and without whom the clinical aspects of this work would not have been possible to pursue and investigate. For their technical assistance and advice, I thank Dr. John Garret, Carol Aspinwall, Lyddia Ruch-Doll, Taylor Hartung, Mary Pierce, Emily Ferris, Jessica Michnik-Rubenstein, and Guy Starbuck. I thank Stephanie Whitehorse, Jeanine Burmania, and Abram Becker for helping to get the technical developments of this work closer to peoples' lives.

For their support and encouragement, for the laughter and the crying, I thank my friends Dr. Daniel Seiter, Dr. Joseph Whitehead, Dr. Ian Marsh, Amelie Dolfi, Abby Huff, Zach Lipp, Victor Estima, Dr. Vinicius Helfstein, Felipe Faleco, and Matheus Pupo.

Finally, I thank my family. My parents, Magaly Santoro, and José Luis Fernandes, for my life and for their unconditional love and support. And my brother, João Marcos Santoro Fernandes, for being my rock. This dissertation is the fruit of your work as much as it is of mine.

# TABLE OF CONTENTS

# LIST OF ABBREVIATIONS AND ACRONYMS

| | |
|---|---|
| ADR | Articulated deformable registration |
| ARR | Articulated rigid registration |
| AUC | Area under the curve |
| CL | Corresponding lesion |
| CNN | Convolutional neural network |
| CT | Computerized tomography |
| DL | Disappearing lesion |
| FFD | Free-form deformation |
| FOV | Field of view |
| HR | Hazard ratio |
| IBI | Inflammation-based index |
| mCRPC | Metastatic castrate resistant prostate cancer |
| MIP | Maximum intensity projection |
| MLR | Multivariable linear regression |
| MM | Metastatic melanoma |
| NET | Neuroendocrine tumor |
| NL | New lesion |
| OS | Overall survival |
| PET | Positron emission tomography |
| PFS | Progression free survival |
| PL | Persisting lesion |
| PPV | Positive predictive value (precision) |
| PRRT | Peptide receptor radiation therapy |
| PSA | Prostate specific antigen |
| RF | Random forest |
| ROC | Receiver operating curve |
| RPCCC | Roswell Park comprehensive cancer center |
| SBRT | Stereotactic body radiation therapy |
| SRI | Somatostatin receptor imaging |
| SSR | Somatostatin receptor |
| SUV | Standardized uptake value |
| UWCCC | University of Wisconsin Carbone cancer center |
| WDR | Whole-body deformable registration |
| WRR | Whole-body rigid registration |

# LIST OF FIGURES

# LIST OF TABLES

# ABSTRACT

Cancer treatment response assessment is particularly challenging in the setting of metastatic cancer in part because patients can present with hundreds of metastatic lesions through their body, which often respond heterogeneously to treatment. The current standard in clinical practice to reach a patient-level disease assessment is a partial response evaluation of a subset of lesions that is grouped and analyzed jointly. This suboptimal practice disregards heterogeneity in lesion-level response as it averages the lesion response, in addition to assessing only a fraction of the lesions. To achieve a more complete understanding of cancer response to treatment, comprehensive, lesion-level, and longitudinal disease assessment is necessary, but precluded by requiring manual matching of corresponding lesions, which is a tedious, subjective, and error-prone task. Therefore, the main goal of the present dissertation is **to develop an automated, comprehensive, lesion-level, longitudinal evaluation of metastatic cancer to enable enhanced treatment response assessment, and to evaluate the improvement in treatment response prediction both at the patient- and at the lesion-level.**

To that end, we have developed and validated a lesion-tracking methodology that allows longitudinal, lesion-level information to be extracted from all the identifiable cancerous lesions in a patient across time. We then investigated the use of such information to build models of patient- and lesion-level treatment response prediction. These predictions were investigated for metastatic neuroendocrine tumor (mNET) patients treated with peptide receptor radiation therapy (PRRT), though the methodology was validated broadly across cancer and treatment types. The main reason for choosing mNETs was the newly introduced PRRT treatment, which is not well understood, especially its widely variable response across patients and lesions.

The developed methodology had high accuracies for lesion matching (0.98) and tracking (0.90), thus enabling automated disease response assessment. When used for treatment response prediction, we showed that the comprehensive, lesion-level, longitudinal information can predict mNET lesions that resist treatment and exhibit a long-term persistence (prediction accuracy of 0.78). On the patient-level, we found that utilizing the information provided by the developed methodology resulted in superior treatment response prediction than previously established methods. We obtained a significant prediction of progression-free survival (PFS) ($p \ll 0.001$) and a stratification between good and poor responders to PRRT better than established approaches (0.88 vs. 0.65 AUC).

In summary, this dissertation is the first work to show that the use of comprehensive, lesion-level, and longitudinal disease information is possible for various types of metastatic solid cancers via an automated methodology. Furthermore, this information was shown to improve the predictability of the treatment response of mNET patients and their individual lesions to PRRT. The potential impact of improved lesion- and patient-level treatment response prediction in the management of cancer patients is promising. Potential uses include: a lesion selection framework for subtotal targeted lesion ablation, inform the selection of patients for clinical trials, optimize treatment monitoring frequency, optimize time of treatment change, and support decisions for combination of different treatments. The introduced treatment response prediction approach has the ultimate intention of meaningfully contributing towards prolonged and higher quality-of-life for metastatic cancer patients.

# 1 INTRODUCTION

## 1.1 METASTATIC CANCER RESPONSE HETEROGENEITY

Metastasis is the leading cause of cancer-related mortality (Chaffer & Weinberg, 2011). Although cancer patients have treatment options, the treatment response among lesions can be substantially different. Some lesions might respond to treatment while other lesions continue to progress. This receives the name of response heterogeneity and poses a great challenge to cancer management in metastatic patients because the behavior of one lesion (or even a few) is likely unrepresentative of the global response. Treatment response heterogeneity is frequent in metastatic cancer and has been observed in various cancer types, like breast cancer (Huyge et al., 2010), melanoma (Carlino et al., 2013), colorectal cancer (Hendlisz et al., 2015; Russo et al., 2016; Van Helden et al., 2016), renal cell carcinoma (Crusz et al., 2016), prostate cancer (Harmon et al., 2016; Kyriakopoulos et al., 2020), non-small cell lung cancer (Dong et al., 2017; Humbert et al., 2020; Y. Lee et al., 2014; Tazdait et al., 2018; Tozuka et al., 2020), and neuroendocrine tumors (Perk et al., 2020). The heterogenous response to treatment is caused by selective treatment resistance, which can be intrinsic or acquired. Intrinsic resistance exists in a treatment naïve setting and is caused by pre-existing random genetic mutations (Maitland, 2021), while acquired resistance can be caused by various factors like selective expansion of pre-existing subclonal cell populations under therapeutic selective pressure (Dagogo-Jack & Shaw, 2018), compensatory increase in survival pathways, or the downregulation of antigens to escape immunological response. An interplay between resistance types can also occur. For example, it has been observed that intrinsic resistance can lead to a higher degree of acquired resistance in metastatic prostate cancer patients (Kyriakopoulos et al., 2020). Importantly, both modes of resistance can lead to clinical progression. Intrinsic resistance lesions can continue to progress, whereas initially responding lesions can later progress. Thus, identifying and targeting

intrinsic resistant lesions, as well as identifying and targeting acquired resistant ones can lead to a delay of clinical progression.

Patients with advanced metastatic cancer are usually treated with systemic therapies that are expected to halt the proliferation of the cancer cells present in every lesion. However, different effects across the various lesions can be observed due to selective treatment resistance. This commonly leads to a temporally limited management of metastatic disease, eventually reaching a failure point, defined as the time to progression (TTP) (Strosberg et al., 2017). Time to progression can be determined by patient-level metrics, such as prostate-specific antigen (PSA) level in prostate cancer, and by imaging criteria such as the RECIST or PERCIST criteria (Eisenhauer et al., 2009; Therasse et al., 2000; Wahl et al., 2009). In clinical practice, when disease progression is clinically determined, it is common to discontinue the systemic treatment. However, for metastatic cancer patients, the determination of patient-level disease progression could be driven by the progression of only one or a few resistant lesions (oligoresistant), while many other lesions are well controlled by systemic therapy. For example, Kyriakopoulos et al. observed that, in metastatic prostate cancer patients, some lesions continue to respond even after progression time has been determined (Kyriakopoulos et al., 2020). This state where a few lesions resist therapy is defined as the oligoprogressive disease state (Palma et al., 2014). In these cases, stopping systemic therapy may not be the right course of action. Importantly, response heterogeneity information is crucial to make an optimal treatment stopping or modification decision, and this information is rarely available. The only way to fully understand the heterogeneity in response, and thus completely understand the metastatic cancer response to treatment, is to acquire spatiotemporally resolved information by tracking *individual* lesions (lesion-level) and *all* the lesions (comprehensive) in a patient across the treatment and post-treatment periods (longitudinally). Medical imaging is the gold-standard method to provide such spatiotemporal resolution.

## 1.2 Medical Imaging

The spatial resolution capabilities confer medical imaging with the capacity to capture different quantitative and qualitative characteristics of each disease site. This allows for the identification of inter-lesion treatment response heterogeneity (Humbert et al., 2020). The increasing role of medical imaging in treatment response assessment is causing the heterogenous characteristic of cancer response to be increasingly noted and investigated (Humbert & Chardin, 2020). Medical imaging can provide anatomical and molecular information, which are complementary in nature. Anatomical Information is a proven efficacious approach to assess patient response to cancer treatments (Curran et al., 2006). It is invaluable since treatment agents that cause the shrinkage of solid tumors are correlated with improvement in overall survival (Eisenhauer et al., 2009). Anatomical-based treatment response guidelines are well established via the RECIST criteria (Eisenhauer et al., 2009). However, despite being fundamental, RECIST is not all-encompassing since some cancer types are characterized by slow growing lesions, for which anatomical changes are preceded by functional changes, for example metastatic neuroendocrine tumors (mNET) (De Mestier et al., 2014). Especially in these cases, molecular information captured by imaging modalities such as PET/CT can play an important role in response assessment. PET/CT imaging captures molecular information in the form of standardized uptake values (SUV) which quantifies underlying biological processes and can be used to detect metastatic lesions and monitor their progression. The biological processes observed are dependent on the biological molecule used in the tracer in association with a radionuclide. For example, $^{18}$F-FDG is the most common PET tracer used and is a marker of glucose uptake, or cell metabolism (Fletcher et al., 2008). The PERCIST criteria are guidelines for evaluating tumor response using $^{18}$F-FDG PET/CT (Wahl et al., 2009). However, all response evaluation criteria (e.g., RECIST and PERCIST) are limited in the amount of information they utilize to perform treatment response evaluation. Furthermore, for tracers introduced more recently, such as $^{68}$Ga-DOTATATE, guidelines for response evaluation criteria are incipient. $^{68}$Ga-DOTATATE is a marker of peptide (somatostatin) expression on the cell membrane, which is overexpressed in well-differentiated neuroendocrine tumors (Kwekkeboom et al., 2010). Molecular

information is becoming increasingly important for lesion identification and treatment monitoring (Baum et al., 2010; Scarpelli et al., 2018; Vanderhoek et al., 2013; Weissleder, 2006).

Beyond the spatially rich nature of molecular imaging, the acquisition of multiple scans across time adds another dimension of information: the temporal. Longitudinal variations in PET tracer uptake can be representative of lesion response to treatment (Wahl et al., 2009). In the span of their disease management, cancer patients can receive multiple scans as part of diagnosis, staging, and treatment monitoring. Longitudinal imaging response assessment of cancer has been identified as a predictor of long-term treatment outcome and adverse events in various settings (Haug et al., 2010; Hribernik et al., 2022). When put together, the spatial and temporal information provided by molecular images has great potential to support clinical patient management through improved treatment response assessment. However, this potential is commonly underutilized.

Frequently, clinical response assessment reduces the rich heterogeneous lesion-level image information to patient-level information either by summarizing all the lesions into a single score (Gallamini et al., 2014; Larson, 1999) or by evaluating a subset of lesions (Eisenhauer et al., 2009; Wahl et al., 2009) (e.g. RECIST or PERCIST). The labor-intensive aspect of the manual task of extracting the lesion-level response information limits the clinical and investigational use of such information. Late-stage cancer patients can present with hundreds of lesions, hence individually evaluating each lesion's temporal evolution (i.e. establishing lesion tracks) is complex, tedious, and error prone (Schwartz et al., 2003). Furthermore, these patients' metastases can be spread through their whole-body, making the lesion-level disease response assessment even more challenging. To unleash the full potential of the cancer treatment response information contained in medical imaging, the manual aspect of the analysis needs to be overcome. For that, the automation of the image analysis tools that allow for the comprehensive extraction of lesion-level response information is necessary.

## 1.3 Comprehensive, Lesion-level, Longitudinal Image Analysis

The image-based lesion-level quantification of disease response relies on a few major image analysis steps: (1) lesion identification and segmentation, (2) longitudinal lesion matching, i.e. establishing longitudinal lesion correspondence from one imaging time-point to the next, (3) quantification of lesion characteristics, and (4) establishing changes of lesion characteristics over time. Lesion identification and segmentation has been the object of much investigation. The pioneer methods have explored optimal threshold limits to perform lesion segmentation (Bosc et al., 2003; Otsu, 1979; A. J. Weisman et al., 2020b). Recently, with the advent of convolutional neural networks (CNN), new and more reliable methods for automated lesion segmentation have emerged (Isensee et al., 2021; Kamnitsas et al., 2017; Schott et al., 2023; A. J. Weisman et al., 2020a). These recent advances have largely fulfilled the necessity for automated lesion identification and segmentation. Therefore, they are not the focus of this work. Instead, we draw on previous works of our research group for these important steps (Perk et al., 2018; Schott et al., 2023; A. Weisman et al., 2022; A. J. Weisman et al., 2020a).

Longitudinal lesion matching, on the other hand, still requires an automated solution that enables its widespread use in clinical practice and academic investigation. The main task of a lesion-matching automated solution is to establish the correct intra-patient correspondence between lesions identified in multiple images acquired at different time-points. Some of the challenges associated with this task are (1) the merging and splitting of lesions that can occur because of treatment response or disease progression, (2) considering multiple imaging modalities used in different time-points, (3) guaranteeing that the lesion-matching performance is satisfactory throughout the entire patient anatomy, and (4) devising a way to optimize the global lesion correspondence decision in the face of multiple concurrent possibilities for matching. If lesion matching automation is achieved, lesion-level response can be assessed and inter-lesion heterogeneity can be quantified, providing potentially valuable information for improved cancer management. Examples of the clinical impact of the assessment of heterogeneous response to treatment include treatment response prediction (Humbert &

Chardin, 2020), patient selection for combined locoregional and systemic therapies (Lyu et al., 2020), and serving as guiding information for lesion biopsies, improving biopsy success rates (Harmon et al., 2016). The development of image analysis tools that enable lesion matching is crucial for automation of comprehensive, lesion-level, longitudinal information extraction, which could lead to improved image biomarkers

## 1.4 Comprehensive, Lesion-level, Longitudinal Image Biomarkers

Image biomarkers are defined as “an objective characteristic derived from an in vivo image measured on a ratio or interval scale as indicators of normal biological processes, pathogenic processes, or a response to a therapeutic intervention" (Kessler et al., 2015). The quantitative biomarkers in PET images commonly consist of statistics calculated from the SUV distributions of volumes of interest (VOI) (Figure 1). One of the most common image biomarkers in PET is $SUV_{max}$, consisting of the maximum value of the SUV distribution in a VOI (e.g.: a lesion or an organ at risk); other commonly used statistics are the mean ($SUV_{mean}$) and the total (integrated distribution - $SUV_{total}$). Comprehensive, lesion-level, longitudinal image analysis can enrich the level of information contained in image biomarkers. Making these images biomarkers more informative can have an impact on treatment response assessment and prediction, since image biomarkers have been shown to be predictive of clinical endpoints such as overall survival and progression-free survival, which are correlated to clinical outcomes.

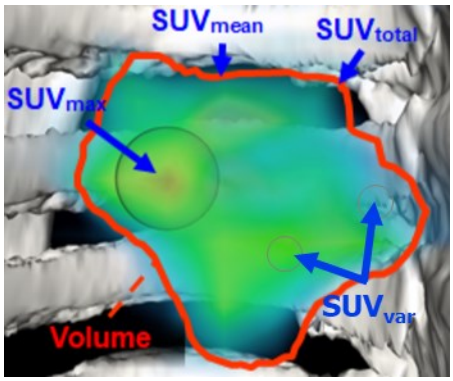


*Figure 1 – examples of image biomarkers extracted from a PET image volume of interest. Adapted from Huff* (Huff, 2022).

Various cancers have shown strong associations between $SUV_{max}$, tumor stage, and prognosis (Al-Taan et al., 2014; H.H. et al., 2014; Namura et al., 2010; Pan et al., 2009; Park & Lee, 2014; Walentowicz-Sadlecka et al., 2014). For example, in lung cancer, many studies have demonstrated that $^{18}$F-FDG PET avidity is associated with increased cellular abnormality (poor tumor biology) and increased risk for recurrence (Melloni et al., 2013). $SUV_{max}$ measured longitudinally on multiple FDG PET/CT scans acquired throughout treatment showed promise in identifying good responders among lung cancer patients, as measured by 80% decreases in $SUV_{max}$ (Cerfolio et al., 2004). Despite these promising results, $^{18}$F-FDG $SUV_{max}$ consistently fails to represent an independent predictor of overall survival in lung cancers (Agarwal et al., 2010). This could be due to the limited nature of the used biomarkers (non-comprehensive, non-longitudinal). Furthermore, it is now accepted that limiting FDG biomarkers to a single measurement (e.g., $SUV_{max}$ ) does not account for the wealth of information contained within the entire tumor volume (Lambin et al., 2012). Another example is metastatic prostate cancer, where studies have demonstrated that both the number of lesions identified by physicians and corresponding uptake on $^{18}$F-NaF PET/CT scans during or after treatment show correlation with clinical outcomes (Apolo et al., 2016; Yu et al., 2015). In a small cohort of patients undergoing chemotherapy, changes in patient-level $SUV_{max}$ on $^{18}$F-NaF PET/CT uptake showed modest correlation with progression-free survival at 12 weeks (Yu et al., 2015). The modest correlation found is possibly limited by the incomplete utilization of image biomarkers. Furthermore, in metastatic neuroendocrine tumor patients, PET-derived image biomarkers have been shown to be correlated to

progression-free survival and overall survival. In a recent systematic review of 39 studies, Lee et al concluded that PET-based biomarkers can impact treatment decision-making and may predict PRRT response in patients with NETs (O. N. Y. Lee et al., 2022). Additionally, the authors issued a call for further research on the topic: "more robust studies should be conducted to better elucidate the prognostic role of somatostatin receptor PET/CT in optimizing treatment for clinical outcome".

## 1.5 Metastatic Neuroendocrine Tumor

There are some characteristics that make comprehensive, lesion-level, longitudinal response assessment particularly useful for metastatic cancers. *First*, showing a large and still poorly understood variation in patient-level response to a systemic therapy. *Second*, showing varied individual lesion response to a systemic treatment (lesion response heterogeneity). *Third*, showing metastases that spread through the whole body. *Fourth*, showing specific pathological consequences associated to lesions that do not respond to therapy. *Finally*, showing disease evolution characteristics that challenge the conventional treatment response assessment (i.e., RECIST or PERCIST). Neuroendocrine tumors are heterogeneous cancers that originate in the diffuse neuroendocrine system and that present all these characteristics[*]. *First*, the recently introduced peptide receptor radiation therapy (PRRT) shows wide variation in patient-level response (Liberini et al., 2020). The introduction of PRRT has significantly improved the standard of care for mNET patients (Kwekkeboom et al., 2008; Strosberg et al., 2017). However, it is well known that NET patients benefit differently from PRRT. For example, Werner et al. reported that in a cohort of 141 NET patients undergoing PRRT, the maximal progression free survival was 85 months while the minimal was 4 days (Werner et al., 2017). *Second*, it has been observed that, when receiving PRRT, the individual lesion response to treatment is varied (lesion response heterogeneity)

[*] Despite the great applicability of the lesion matching methodology to mNET patients, other cancer types (e.g., metastatic melanoma) present similar characteristics. Therefore, care was taken in the lesion matching methodology development to make it broadly applicable. In its development, we have used images from patients with multiple types of metastatic solid cancers.

(Krenning et al., 2005; O'Connor et al., 2017; Roll et al., 2021). *Third*, the most common type of NET is originating in the midgut and the pancreas, but it can also originate in lung tissue and the pituitary gland (Kulke et al., 2012), and it can metastasize to the whole body, with sometimes hundreds of lesions distributed through various types of tissues (e.g.: bone, liver, lymph nodes). *Fourth*, almost half (about 40%) of NETs secrete hormones, causing carcinoid syndrome and greatly reducing patients' quality of life. *Finally*, most NETs are slow growing (Ahmed, 2020), which brings challenges to the classical anatomical treatment response assessment (i.e.: RECIST). Furthermore, despite being a relatively rare cancer, mNET incidence in the United Stated increased from 1 in 100,000 in 1973 to 7 in 100,000 in 2012 (Ahmed, 2020). It is believed that improved imaging technology and improved recognition of neuroendocrine diseases are the reason for the perceived frequency increase (Cives & Strosberg, 2018). The imaging of mNET patients with somatostatin receptor imaging (SRI) (e.g., $^{68}$Ga-DOTATATE PET/CT) is a clinical standard to select patients for PRRT. Clinical guidelines state that the uptake of all lesions must be higher than the background uptake of the liver at a baseline SRI to start PRRT, i.e.: Krenning score of at least grade two (Becx et al., 2022). When performed longitudinally (before and after treatment), SRI depicts the early lesion-level response to PRRT and provides a spatiotemporal complete portrait of the early disease response (Sahani et al., 2013). Using SRI images, biomarkers have been studied for treatment response assessment and prediction in mNET patients. However, the potential of comprehensive, lesion-level, longitudinal PET biomarkers remains unexplored.

## 1.6 Treatment Response Prediction

In some studies, PET imaging has been shown to be an early predictor of long-term response to treatment (Campana et al., 2010; Haug et al., 2010). Motivated by these results, various studies have assessed the value of SRI for predicting PRRT response of NETs (O. N. Y. Lee et al., 2022). However, the image information available to explore the treatment response prediction has always been limited, which hinders the predictive value of the images and precludes the exploration of lesion-level response prediction. On the lesion-level, comprehensive

response assessment enables investigation on the prediction of resistant lesions. These lesions could be candidates for receiving locoregional therapies in addition to the systemic therapy regimen. Locoregional therapies have been shown to improve patient quality of life and increase survival in patients with oligometastatic disease (Akyildiz et al., 2010; Elias et al., 2009; Hafez & Gettinger, 2020; Mohan et al., 2015). Oligometastatic is the state where patients have few metastases, commonly defined as less than five, and all metastases are limited to one organ (Palma et al., 2014). In contrast to the oligometastatic stage, delivering locoregional treatment to widely metastatic cancer patients, as in the case of most mNET patients, presents a greater clinical challenge due to the elevated number of lesions. Invasive interventions performed on tens of lesions can be accompanied by high rates of morbidities and complications (Elias et al., 2009). However, if provided with comprehensive, lesion-level response assessment, clinicians could perform subtotal locoregional therapies, selecting only the resistant lesions to receive the locoregional treatments. The use of locoregional therapies in mNET patients has achieved some success, for example, stereotactic body radiation therapy (SBRT) has been used with curative intent (Jo & Yeo, 2019; Sharabi et al., 2017). The benefits observed in oligometastatic patients who received locoregional ablation could then potentially be transferred to oligoprogressive patients receiving systemic therapies, such as mNET patients receiving PRRT. In this way, selective locoregional intervention to resistant lesions could delay disease progression and improve patient quality of life. The potential benefits would have to be tested in phase II clinical trials; however, studies of feasibility are needed prior to a prospective trial.

Going beyond the lesion-level response prediction, we hypothesize that comprehensive, lesion-level, longitudinal image biomarkers can improve the patient-level prediction of treatment outcomes because of capturing the heterogeneous (inter-lesion) disease response. Various studies have assessed the value of SRI for predicting therapy response of NETs (Figure 2), however they have always been limited in various ways. Efforts into investigating the SRI predictive power for PRRT in NETs started over a decade ago when Haug et al and Gabriel et al investigated the predictive power of the early functional response of a single lesion (Gabriel et al.,

2009; Haug et al., 2010; Öksüz et al., 2014). The SUV of a limited subset of lesions at baseline SRI has also been investigated as a predictor of PRRT in NET patients (Ambrosini et al., 2015; Atkinson et al., 2021; Campana et al., 2010; Carlsen et al., 2020; Graf et al., 2020; Kratochwil et al., 2015; Soydal et al., 2016; Teker & Elboga, 2021; Werner et al., 2017, 2019; Zwirtz et al., 2022). While these works have expanded the number of lesions observed, they forfeited the valuable longitudinal early functional response information. This information has been taken into account by a smaller number of works (Huizing et al., 2020; Ortega et al., 2021; Sharma et al., 2019; Urso et al., 2023; Weber et al., 2020). These investigations, however, remain limited as to the extent of disease being used to explore the SRI predictive potential since a non-comprehensive subset of lesions were quantified and analyzed. PRRT is indicated for advanced, well-differentiated, normally widely metastatic mNET patients hence these patients present with tens, sometimes hundreds of metastases throughout their bodies and limiting the number of lesions from which the biomarkers are used for prediction implies a significant underutilization of the available information

More recently, with the advent of methodologies to automatically or semi-automatically segment all lesions in a PET/CT image (Carlsen et al., 2021; Wehrend et al., 2021), the complete set of lesions in the SRIs could be segmented and quantified. Taking advantage of such technologies, Carlsen et al. investigated the predictive power of SRI uptake of the complete set of lesions (Carlsen et al., 2021). However, this study did not take longitudinal information into consideration. Recent works investigated the predictive power of SRI using both the complete set of lesions and longitudinal early response information (Durmo et al., 2022; Ohlendorf et al., 2020; Opalińska et al., 2022; E. Pauwels et al., 2020). However, these studies evaluated early variation of patient features (patient-level response), and the individual early variation of lesion features (lesion-level response) was not calculated. Both approaches to variation calculation are shown in equations (*1*) and (*2*).

$$\text{Patient-level response:} \quad \left[\sum_i iSUV_{j,post-T}\right] - \left[\sum_i iSUV_{j,baseline}\right] \tag{1}$$

$$\text{Lesion-level response: } \sum_{i}\left[iSUV_{j,post-T} - iSUV_{j,baseline}\right] \quad (2)$$

where $i$ indexes the lesions and $j$ indexes the measured feature (e.g.: max, mean, etc.). Therefore, until this moment, no work has performed the SRI-based treatment response prediction of mNET patients using comprehensive, lesion-level, and longitudinal PET biomarkers.

To conclude, a lesion-level understanding of the spatiotemporal patterns of disease progression can provide a framework for novel local and systemic therapy combinations, but also has the potential to enhance patient-level response prediction. Therefore, this dissertation aims on assessing the added value of using comprehensive, lesion-level, longitudinal information for mNET patients in two scenarios: identification of resistant lesions (Lubner et al., 2013; Roth et al., 2020) and prediction of patient-level treatment outcomes (Gayed et al., 2004; Schwarz et al., 2005).

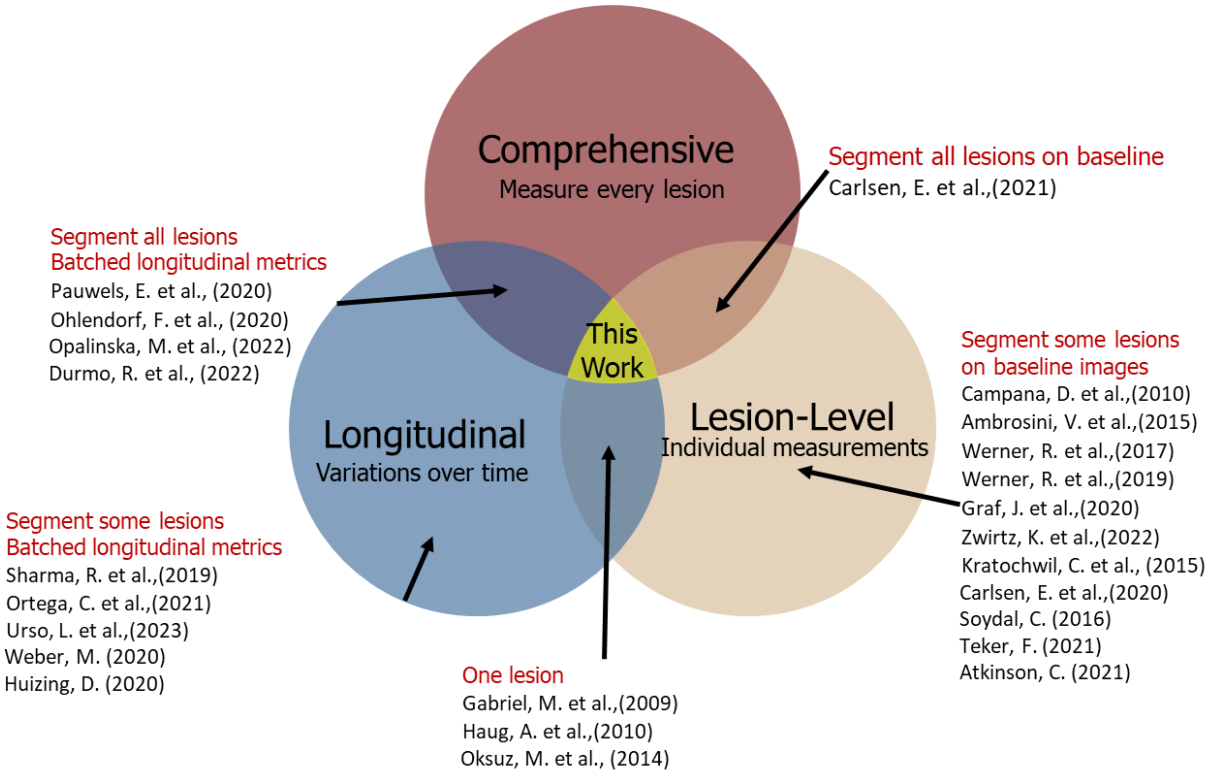


*Figure 2 – Venn diagram of previous works and their characteristics of SRI information usage for PRRT response prediction.*

## 1.7 Aims

We hypothesize that **comprehensive, lesion-level, and longitudinal imaging biomarkers will lead to improved treatment response prediction because of capturing disease response heterogeneity**. Therefore, the main goal of the presented thesis was **to develop an automated, comprehensive, lesion-level, longitudinal evaluation of metastatic cancer to enable enhanced treatment response assessment, and to evaluate the improvement in treatment response prediction both at the patient- and at the lesion-level.** To this end, we have the following aims (Figure 3).

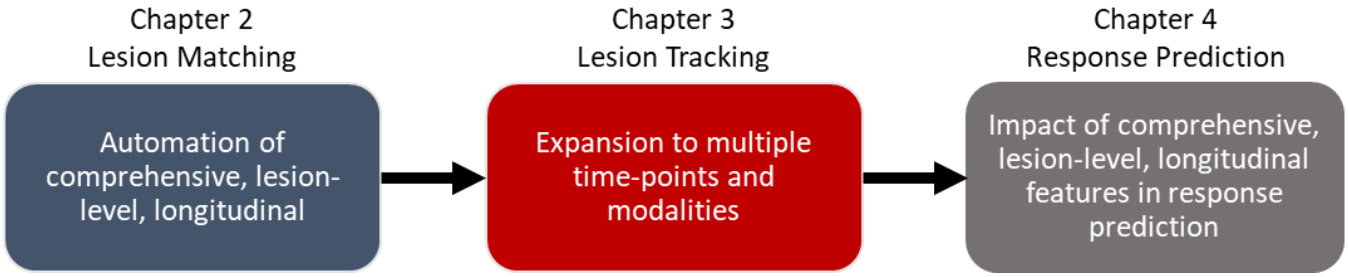


*Figure 3 – Schematic representation of the dissertation's organization*

Aim 1: To develop a method for automated comprehensive, lesion-level, longitudinal matching of disease. Manual disease response assessment between consecutive imaging time-points is clinically infeasible on a lesion-level but is paramount to the complete utilization of the information available in medical images. There is, therefore, a clear need to automate the quantification of such treatment response assessment metrics. The description of the method's development is described in Chapter 2.

Aim 2: To expand the developed lesion matching methodology to allow for the assessment of multiple imaging time-points, as opposed to a consecutive pair, and to allow for the analysis of lesions identified in different imaging modalities. Cancer patients are imaged with multiple modalities on various time-points as part of their treatment response assessment. It is therefore paramount that the comprehensive, lesion-level, longitudinal analysis can be used in these scenarios. The description of the method's expansion is described in Chapter 3.

Aim 3: To assess the value of comprehensive, lesion-level, longitudinal disease metrics for treatment outcome prediction on the lesion- and patient-level. On the lesion-level, improved response prediction can inform the use of selective locoregional therapies. On the patient-level it can help inform decisions regarding treatment tradeoffs and explore different treatment options. This investigation described in Chapter 4.

# 2 DEVELOPMENT AND VALIDATION OF A LONGITUDINAL SOFT-TISSUE METASTATIC LESION MATCHING ALGORITHM

In chapter 2, the lesion matching problem is introduced and tackled for a pair of imaging time-points. Material from this chapter is adapted from a manuscript titled "Development and validation of a longitudinal soft-tissue metastatic lesion matching algorithm" published in Physics in Medicine and Biology in 2021 (Santoro-Fernandes et al., 2021). Furthermore, the methodology was patented through the Wisconsin Alumni Research Foundation under the title "system and method for monitoring multiple lesions", number P210212US01.

## 2.1 MOTIVATION

In this chapter we aim to develop a methodology for automated lesion matching across paired images. The matching of individual lesions across multiple time points has been investigated via rigid registration for identification of new cancerous lesions (Moltz et al., 2009) and for treatment response assessment of multiple sclerosis brain lesions (Bosc et al., 2003; Köhler et al., 2019; Shahar & Greenspan, 2005). Going beyond rigid registration, Yip et al introduced a hierarchical articulated registration that enabled treatment response assessment of late-stage prostate cancer bone lesions (Yip et al., 2014; Yip & Jeraj, 2014). Finally, two dimensional (2D) deformable registration was previously explored as a means of refining the rigid registration-based lesion matching of lymphoma lesions (Xu et al., 2011).

The aforementioned studies demonstrated that rigid, articulated, and deformable image registration enable lesion matching in specific organs (e.g. skeleton, brain, liver). However, a complete algorithm that enables matching of metastatic lesions spread across the entire patient anatomy remains unavailable. Rigid registration applicability is limited to certain body-parts (e.g. brain) and fails to account for skeletal movement

and soft-tissue deformations. Deformable registration showed promising results to aid in lesion matching, but the 2D nature of its previous application fell short on exploring the full potential of the deformations.

In this chapter, we introduce a novel algorithm for lesion matching of metastatic lesions spread through the whole-body that is based on several distinctive steps, including 3D deformable registration, lesion dilation, clustering, and linear assignment of lesions. The lesion matching algorithm was tested in a population of cancer patients that received longitudinal PET/CT scans as part of their treatment response assessment.

## 2.2 Algorithm Description

We pose the lesion matching task as a linear assignment problem. Following image registration (Step 1), the matching algorithm includes lesion dilation (Step 2), clustering (Step 3), and linear assignment (Step 4). A schematic description of the algorithm's structure is presented in Figure 4.

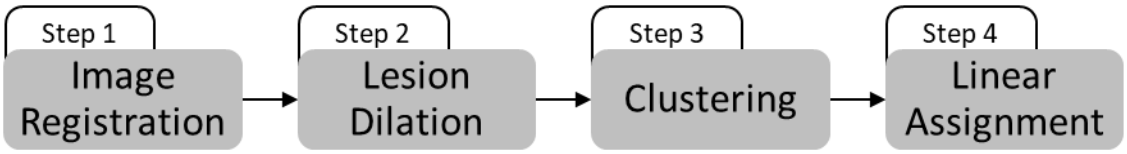


*Figure 4 - schematic description of the lesion matching algorithm*

The algorithm acts on a scan-pair of anatomical medical images, each one having an associated binary lesion mask. The goal of the algorithm is to automatically perform lesion matching decisions based on lesion superposition. A matching decision consists of identifying if a lesion present in one scan has an associated lesion (homologous lesion) in the paired scan. If so, the longitudinal pair of lesions is called single corresponding lesion (CL). Lesions that have no corresponding lesion in the paired scan are either new lesions (NL) or disappearing lesions (DL). In each scan-pair, one matching decision has to be made for each unique lesion and the number of unique lesions per scan-pair is defined as the sum of CLs, NLs, and DLs. The concepts of scan-pairs, unique lesions, matching decisions, corresponding lesions, new lesions, and disappearing lesions are illustrated in Figure 5.

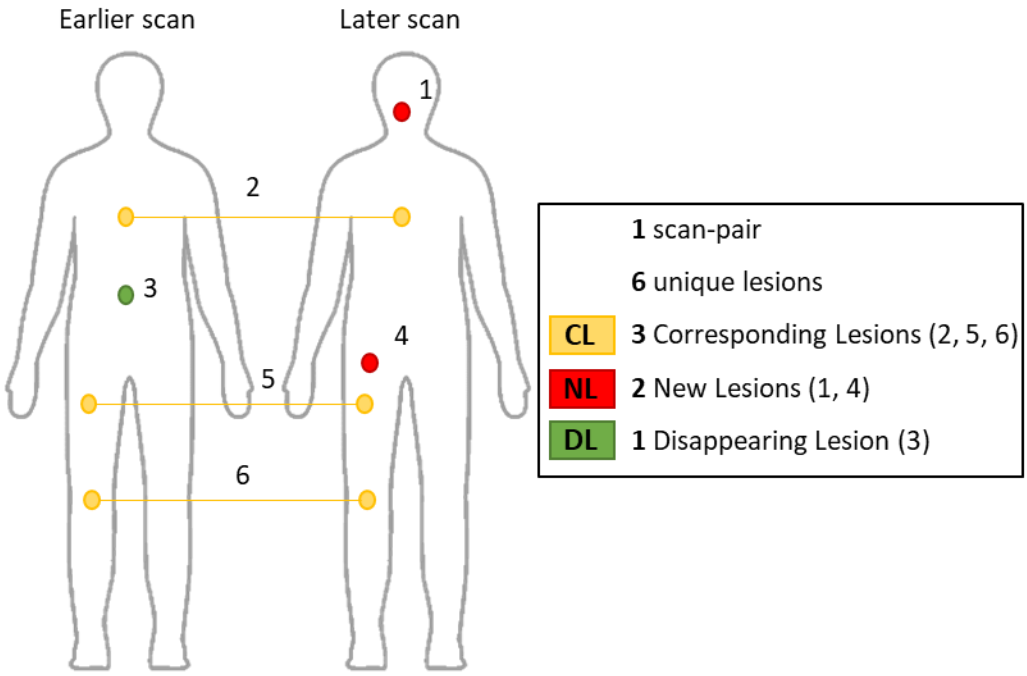


*Figure 5 – Illustration of types of unique lesions per scan-pair: corresponding lesions (CL), new lesions (NL), and disappearing lesions (DL).*

In the **first step**, 3D deformable image registration is used to superimpose anatomically homologous regions of a scan-pair. The deformation vector field is estimated by minimization of the sum of square differences between voxel intensities in the moving and the fixed anatomical medical images, respectively the later and the earlier scans. Deformable registration is done using free-form deformation with 3$^{rd}$ order B-splines interpolation and hierarchical control grid (Rueckert et al., 1999). This technique was chosen for its wide acceptance in the medical image registration field, its local support, and relatively simple and fast calculations (Sotiras et al., 2014). The hierarchical control grid approach is also well established and is known to help in preserving the smoothness of the registration (S. Lee et al., 1996; Mattes et al., 2003; Rueckert et al., 1999).

The calculated transformation field $T$ is then applied to the binary lesion mask of the later scan in the scan-pair,

$$L_T = T(\mathrm{L}) \qquad (3)$$

where $L$ is the set of coordinates of the lesions in the binary mask of the later scan and $\mathrm{L_T}$ is this set of coordinates after transformation.

In the **second step**, a morphologic conformal lesion dilation operation ($D$) is applied to the binary lesion masks (on both time points in the scan-pair) to increase the lesion superposition probabilities (Figure 6).

$$L_{T,D} = D(L_T) \tag{4}$$

$$E_D = D(E) \tag{5}$$

where $E$ is the set of coordinates of the lesions in the binary mask of the earlier scan, $E_D$ is the conformal dilation of the set $E$, such that $E \subset E_D$, and $L_{T,D}$ is the conformal dilation of the set $L_T$, such that $L_T \subset L_{T,D}$. Subsequently, the individual lesion components are numbered based on connected components analysis, the superposition of the lesions is evaluated and a matrix $M$ of intersections is automatically constructed as

$$M_{ij} = \left| E_{D_i} \cap L_{T,D_\mathrm{j}} \right| \tag{6}$$

where $L_{T,D_j} \subset L_{T,D}$ is the subset of the coordinates of the j[th] lesion in the later scan, $E_{D_i} \subset E_D$ is the subset of the coordinates of the i[th] lesion in the earlier scan, and $M_{ij}$ is the cardinality of the intersection of these subsets. The matrix $M$ is a square matrix with $N^2$ elements where $N$ is the maximal integer between $N_E$ and $N_L$, which are the number of lesions identified in the earlier and in the later scans.

The algorithm's **third step** involves the clustering of lesions, an important consideration for treatment response assessment, which can occur due to merging or splitting of multiple lesions from one to another imaging time point (Shafiei et al., 2021). Merging is usually a result of disease progression in adjacent lesions. Conversely, lesion splitting is usually caused by one lesion responding to treatment and separating into two or more smaller disjoint lesion components.

After the second step, a lesion may intersect multiple lesions on the paired scan. This may be a spurious intersection with a neighboring lesion (Figure 6C), or may be due to merging or splitting (Figure 6D). We developed a novel lesion clustering algorithm designed specifically to distinguish between these two instances, so that the overall matching algorithm considers sets of components originating from splitting, and sets of components that eventually merge, as individual lesions. A set of lesion components is clustered when:

i. The components intersect to at least one common lesion on the other scan (Figure 7B), considering the dilation introduced on the second step
ii. The distance $\vec{u}$ between the centroids of the components (Figure 7C) is smaller than a distance $\vec{d}$ (Figure 7D)

The distance $\vec{d}$ is defined as the longest chord in the common matching lesion that makes a smaller than 5° angle with the vector $\vec{u}$ connecting the centroids of the components to be clustered (Figure 7), the angular constraint is expressed as

$$\arccos\left(\frac{\vec{u}\cdot\vec{d}}{|\vec{u}|\cdot|\vec{d}|}\right)\leq 5°. \tag{7}$$

Assuming perfect registration, if a set of components originate from a homologous lesion in the paired scan (either merging or splitting), the components should all be spatially contained in the homologous lesion, in which case the condition $|\vec{d}| > |\vec{u}|$ will be satisfied. If registration is not perfect, but the dilation step guarantees that all the components overlap to the homologous lesion, the relationship between the vectors $\vec{u}$ and $\vec{d}$ will still hold. The angular constraint of equation (7) is imposed to consider that a chord $\vec{d}$ in a direction in which the homologous lesion is long might be an overly lenient threshold for evaluating a vector $\vec{u}$ connecting components separated along a different direction, in which the homologous lesion is short. The necessity for an angular constraint is illustrated in *Figure 8*. The 5° tolerance in equation (7) is used to consider that $\vec{d}$ and $\vec{u}$ need not be exactly parallel, but approximately so. The 5° value is an algorithm hyperparameter and different values could be explored depending on the application.

In the **fourth step**, the lesions are renumbered and voxels that pertain to a given lesion are reevaluated based on the clustering decision. A new intersection matrix $M'$ is calculated under the new conditions. A cost matrix $C$ is then defined as $C_{ij} = 1/M'_{ij}$. The lesion matching problem is thus reduced to a linear assignment problem (Jaqaman et al., 2008), which consists of finding an optimal permutation matrix $A$

$$A = \underset{A}{\operatorname{argmin}} \left( \sum_{i=1}^{N} \sum_{j=1}^{N} A_{ij} C_{ij} \right) \tag{8}$$

within the constraints to the permutation matrix $A$

$$\sum_{i=1}^{N} A_{ij} = 0 \; or \; 1 \text{ and } \sum_{j=1}^{N} A_{ij} = 0 \; or \; 1 \tag{9}$$

where summations might be zero if a lesion is superimposed to no other (NLs or DLs). The Munkres assignment algorithm (Munkres, 1957) is used to solve the linear assignment problem. One of the possible mathematical interpretations of the Munkres algorithm is that it finds the optimal permutation matrix $\mathrm{A}$ that minimizes the global sum of costs defined by equation (8) within the constraints imposed by equation (9). Briefly, the Munkres algorithm performs a non-greedy minimization of $C$, which is equivalent to a maximization of the volume of intersection between all lesions in a scan-pair.

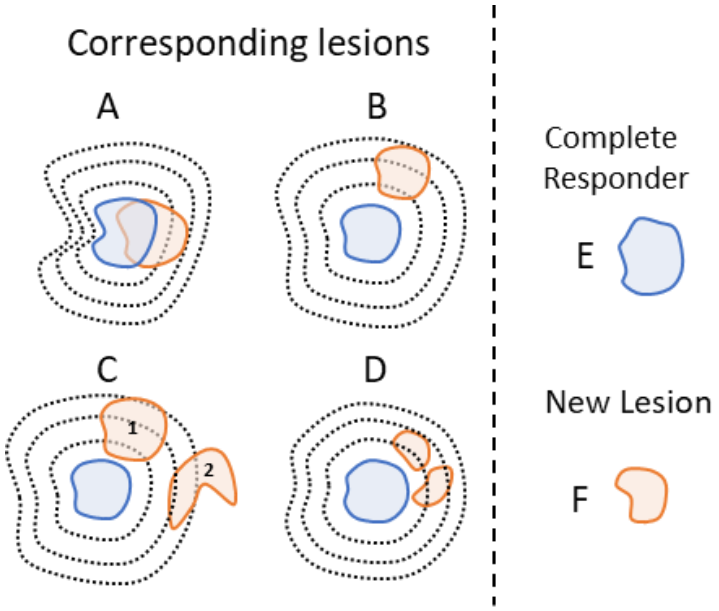


*Figure 6 – Illustration of the conformal dilation role in lesion matching. For clarity, lesion dilation is represented only on one scan of the scan-pair (blue). Lesions (solid contours) and lesions' conformal dilations (black dotted lines) colors indicate lesions from different scans in a scan-pair. (A) Corresponding lesions are matched based on direct superposition. (B) Despite no direct superposition, match is still possible because of the conformal dilation. (C) lesion dilation yields a correct (1) and one incorrect (2) superposition, however, the Munkres algorithm selects the correct match. (D) Two lesions are superimposed to the dilation and are merged and matched. (E) A single disappearing lesion. (F) A single new lesion.*

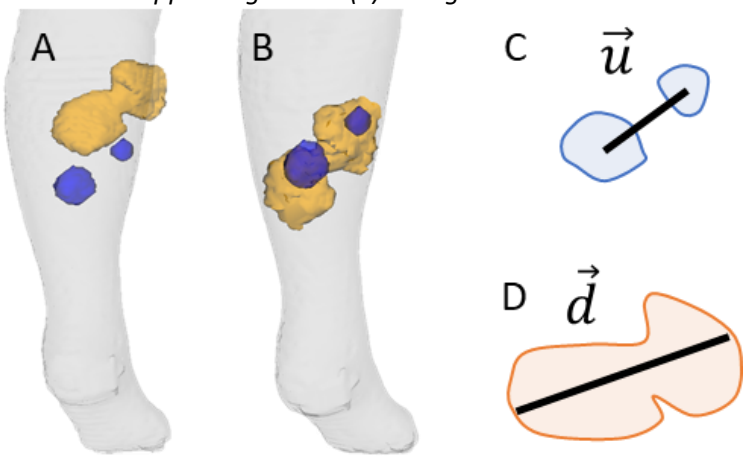


*Figure 7 – Situation illustrating clustering as lesion merging caused by disease progression. The blue masses are from an earlier scan and the orange mass is from a later scan. (A) Before registration, the lesions are not superimposed. (B) After registration, lesions are superimposed. (C) illustrates the vector $\vec{u}$ between centroids and (D) the vector $\vec{d}$. After the clustering, the two blue masses correspond to one single lesion (orange).*

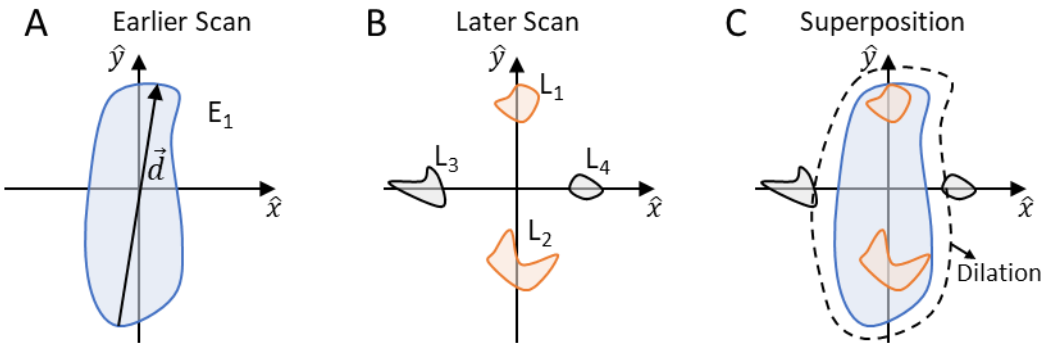


Figure 8 - Illustration of the rationale for using a small angle constraint in the lesion clustering step. If lesion overlap were considered alone, lesions $L_1$, $L_2$, $L_3$, and $L_4$ would cluster since all of them overlap to lesion $E_1$ (considering dilation). Using the constraint $|\vec{u}| < |\vec{d}|$ without an angular constraint would still be insufficient and imply in wrongful clustering of $L_1$, $L_2$, $L_3$, and $L_4$. Using the angular constraint, $L_3$ and $L_4$ are correctly excluded from the cluster.

## 2.3 Algorithm Validation

The developed algorithm was evaluated in 172 PET/CT images of 32 patients with metastatic cancer, all images were acquired under the institutional standard procedures. PET/CT images provide anatomical structure and molecular information of the location of diseased tissue, which makes them suitable for validating the algorithm. The PET/CT images were retrospectively acquired from two independent clinical trials (Table 1). One of the trials contained 19 metastatic melanoma patients, for whom lesion-level disease assessment is particularly challenging because of the plurality of metastatic sites and the elevated number of lesions, which in the most extreme case was of 64 unique lesions spanning from the legs to the head. A second dataset containing 10 different types of cancers was included to evaluate the algorithm for its robustness regarding disease type.

In the validation dataset, a scan-pair was defined as two longitudinal and sequential PET/CT scans of the same patient. Therefore, if $n$ is the number of scans that a patient received, $n-1$ is the number of scan-pairs; e.g. a patient that received four scans $(s_1, s_2, s_3, s_4)$, will have three scan-pairs $s_1$ to $s_2$, $s_2$ to $s_3$, and $s_3$ to $s_4$. The dataset contained 140 scan-pairs.

*Table 1 - Patient population description*

| # of Patients | Tracer | Scanner | Disease | Treatment | Design |
|---|---|---|---|---|---|
| 19 | $^{18}$F-FDG | GE Discovery IQ<br>GE Discovery 710<br>GE Discovery STE | Metastatic Melanoma | immune checkpoint inhibitor and antiangiogenic drugs | Retrospective |
| 13 | $^{18}$F-FLT | GE Discovery VCT | Upper tract urothelial, rectal, breast, lung (non-small cell), prostate, colon, ovarian, appendiceal, adenoid, and squamous cell | antiangiogenic drug | Prospective (Scarpelli et al., 2016) |

In the validation, CT images served as the anatomical information, while the binary lesion masks were derived from the PET image. Using radiology reports as guidance, cancerous lesions, indicated by elevated $^{18}$F-FDG or $^{18}$F-FLT PET SUV uptake relative to background, were manually contoured, defining thus the binary lesion masks associated to each PET/CT image. The manual contouring started by using an active-contours algorithm to automatically conform the contours to the lesions (Yang et al., 2009), then the contours were reviewed by a nuclear medicine physician and manually corrected when necessary. Reference matching decisions were made by visually analyzing the scan-pairs' images for each patient and preparing a reference lesion matching decision matrix for each scan-pair. The merging or splitting of lesions over time was accounted for in the reference matching decisions. Clustering decisions were made when two or more lesions occupied the same anatomical position within the patient body as a larger lesion in the corresponding scan (Figure 7B).

The main metric used to assess the algorithm performance on the validation dataset was the overall accuracy of lesion matching, determined as the number of correct lesion matching decisions made divided by the total number of matching decisions. Additionally, the accuracy of matching corresponding lesions and the accuracy of identifying vanishing lesions and new lesions were estimated separately.

#### *2.3.1.1 Image Registration Sensitivity Analysis*

To fully understand the impact of image registration on lesion matching, we studied the algorithm performance under four distinct registration methods. Two deformable registration approaches were investigated: (1) Articulated Deformable Registration (ADR) which was preceded by Whole-body Rigid Registration (WRR) and Articulated Rigid Registration (ARR) and (2) Whole-body Deformable Registration (WDR), which was preceded only by WRR (Figure 9). Both deformable registration approaches (WDR and ADR) were done using free-form deformation with 3$^{rd}$ order B-splines interpolation and hierarchical control grid (Rueckert et al., 1999). Additionally, lesion matching was assessed after WRR and ARR alone.

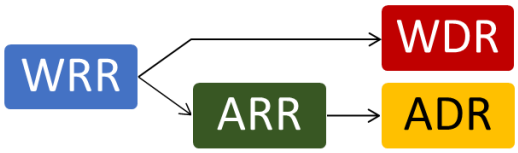


*Figure 9 – Diagram illustrating the approaches to arrive at whole-body and articulated deformable registrations.*

The centroids of corresponding lesions were used as landmarks to analyze the performance of image registration, only lesions whose volume varied less than 50% between scans were used because high volume variation can lead to displacements of the lesion centroids which are unrelated to registration performance. The Wilcoxon signed-rank test was used to compare the centroid distance distributions. Additionally, the impact of the registration methods on lesion matching performance was assessed using the McNemar test to intercompare all the registration methods (0.05 significance level).

#### *2.3.1.2 Lesion Dilation Sensitivity Analysis*

A sensitivity study was conducted to assess dependence of the lesion matching accuracy on the magnitude of the conformal dilation (Step 2) by varying the dilation from 0 to 55 mm in 5 mm increments.

#### *2.3.1.3 Anatomic Distribution Sensitivity Analysis*

The overall lesion matching accuracy evaluation was stratified for the whole body and per individual body parts. The individual body parts considered were head + neck, chest + abdomen, pelvis, upper arms, forearms, hands, thighs, legs, and feet.

#### *2.3.1.4 Lesion Burden Sensitivity Analysis*

The lesion matching accuracy was analyzed on a subset of high burden scan-pairs to investigate if an accuracy drop would be observed in those cases. High burden scan-pairs were defined as those containing 10 or more unique lesions, twice as much as what is recommended for visual human evaluation. In a clinical setting and without automation, individual comprehensive lesion-level response assessment of these scan-pairs is prohibitive. Therefore, satisfactory performance of the algorithm on these scan-pairs is paramount for its possibility of translation into the clinic.

## 2.4 RESULTS

In this study, 736 matching decisions were necessary to analyze all 140 scan-pairs. Based on the reference matching decisions, 298 of the 736 matching decisions (40%) were corresponding lesions and 438 (60%) were new or disappearing lesions. The accuracy of lesion matching is shown in Table 2. These results represent the accuracy with no dilation.

*Table 2 - Accuracy of lesion matching when no dilation is used.*

| | **Accuracy** | | |
|---|---|---|---|
| Registration | Corresponding Lesions N = 298 | New or disappearing lesions N = 438 | Overall N = 736 |
| Whole-body deformable (WDR) | 0.75 | 1.00 | 0.90 |
| Articulated deformable (ADR) | 0.72 | 0.99 | 0.88 |
| Whole-body rigid (WRR) | 0.58 | 0.99 | 0.82 |
| Articulated rigid (ARR) | 0.68 | 1.00 | 0.87 |

#### *2.4.1.1 Image Registration Sensitivity Analysis*

The centroids of 186 corresponding lesions were used as landmark. The landmark distance distributions are shown in Figure 10. All distributions were significantly different ($P < .001$). A visual understanding of how registration influences landmark distance, and thus proximity of corresponding lesions, is shown in Figure 11.

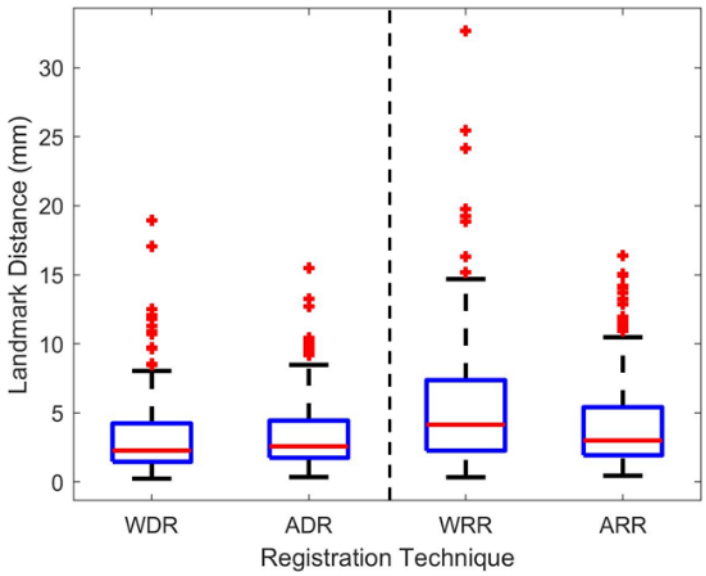


*Figure 10 – Landmark distance distribution of the 186 landmark lesions. Results on the left are for the 3D deformable registration methods: whole-body deformable (WDR) and articulated deformable (ADR). Results on the right are for previous explored registration methods: whole-body rigid (WRR) and articulated rigid (ARR).*

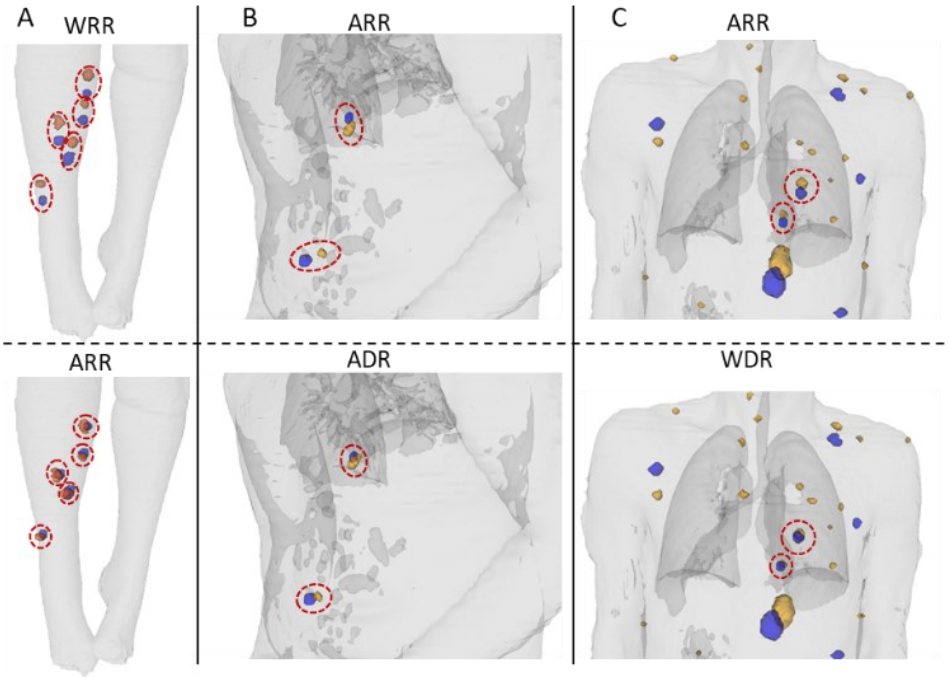


*Figure 11 -Visualization of the impact of different registrations on proximity of corresponding lesions. The CT contour is shown in light gray. Earlier scan lesions are shown in blue and later scan lesions in orange. Red ellipses highlight the corresponding lesions that had visible proximity. Images are shown for whole-body rigid (WRR), articulated rigid (ARR), articulated deformable (ADR), and whole-body deformable (WDR) registrations. Column (A) compares WRR to ARR. Column (B) compares ARR to ADR. And column (C) compares ARR to WDR.*

The McNemar test was used to intercompare the deformable registration approaches' impact on the lesion matching performance. Lesion matching performance using WDR was better than ADR, nevertheless with

marginal statistical significance $(P = .04)$. We also used the McNemar test to compare both deformable approaches to the rigid approaches, WRR and ARR. Comparison of both approaches with WRR yielded statistically significant differences $(P < .001)$. WDR significantly improved lesion matching when compared to ARR ($P < .001$). Improvement observed with ADR was close on the verge of significance ($P = .06$).

#### *2.4.1.2 Lesion Dilation Sensitivity Analysis*

An increase in matching accuracy was observed for corresponding lesions across the dilation range (Figure 12A). By contrast, the accuracy for new or disappearing lesions decreased with lesion dilation, (Figure 12B). The overall accuracy was observed to have an initial increase associated with the CL accuracy trend, followed by a decrease associated with the NL/DL accuracy trend (Figure 12C). The optimal dilation magnitude was determined by the peaks in the overall accuracy curves. The optimal dilations were of 25 mm for all registration methods besides WRR, which had an optimal dilation of 30 mm. The optimal overall accuracy of 0.98 was achieved for WDR at 25 mm dilation.

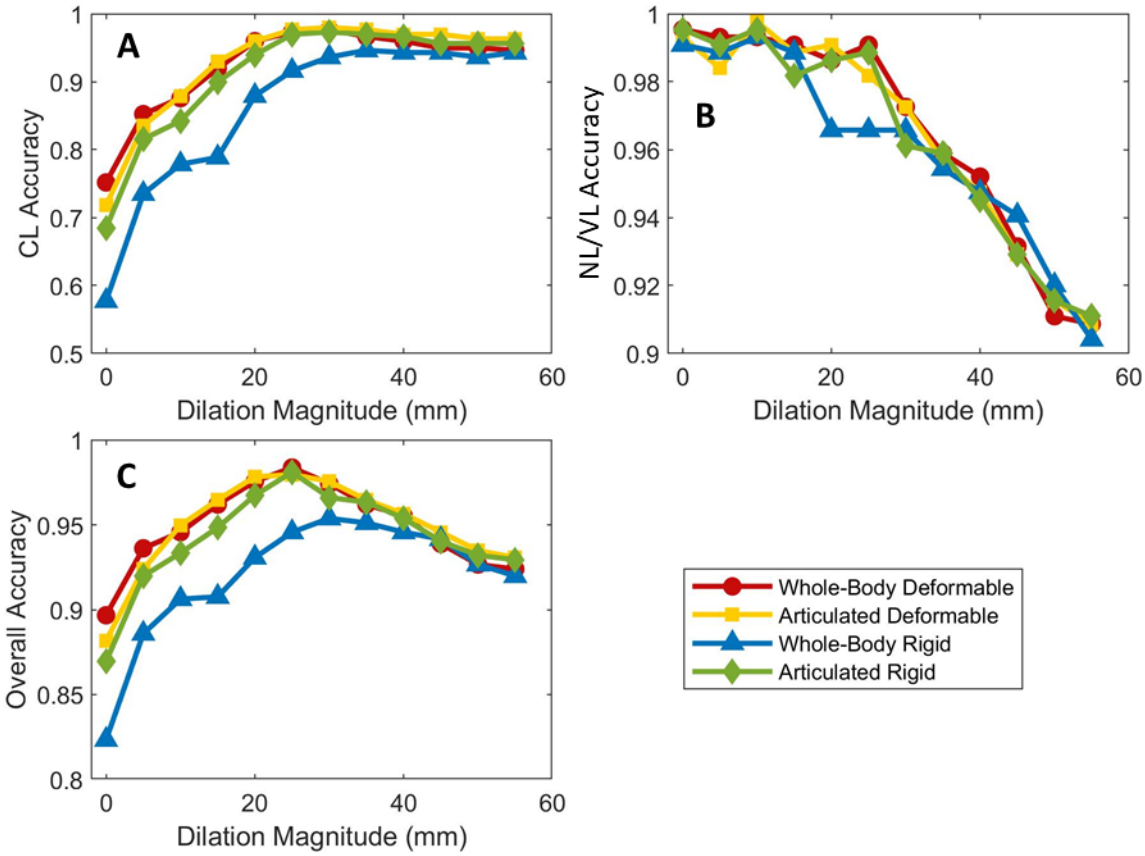


*Figure 12 – Accuracy as a function of dilation magnitude for (A) corresponding lesions (CL), (B) new or disappearing lesions (NL/DL), and (C) all lesions. The markers corresponding to 0 mm of dilation represent the accuracy of the method without lesion dilation.*

### *2.4.1.3 Anatomic Distribution Sensitivity Analysis*

The data in Table 3 are derived from our observation of the 736 lesions present in the patients in our datasets, the location of all unique lesions are shown in Table 3, most of which were the chest+abdomen (52% - 380/736) and in the pelvis (21% - 156/736). The thighs (13% - 99/736) and the legs (8% - 60/736) also presented with relatively high disease densities.

*Table 3 – Number of lesions per body-part*

| Body-part | # of lesions | % | Body-part | # of lesions | % |
|---|---|---|---|---|---|
| Chest+abdomen | 380 | 52% | Feet | 9 | 1.2% |
| Pelvis | 156 | 21% | Forearms | 5 | 0.7% |
| Thighs | 99 | 13% | Arms | 5 | 0.7% |
| Legs | 60 | 8% | Hands | 3 | 0.4% |
| Head & neck | 19 | 3% | | | |

For the four body-parts with highest number of lesions (chest+abdomen, pelvis, thighs, and legs), assessment of the overall accuracy was performed separately. The overall accuracies per body-part, evaluated at the optimal dilations (Figure 13) saw minimal variation across registration methods and body-parts. No accuracy result was smaller than 0.90 (WRR in the Thigs).

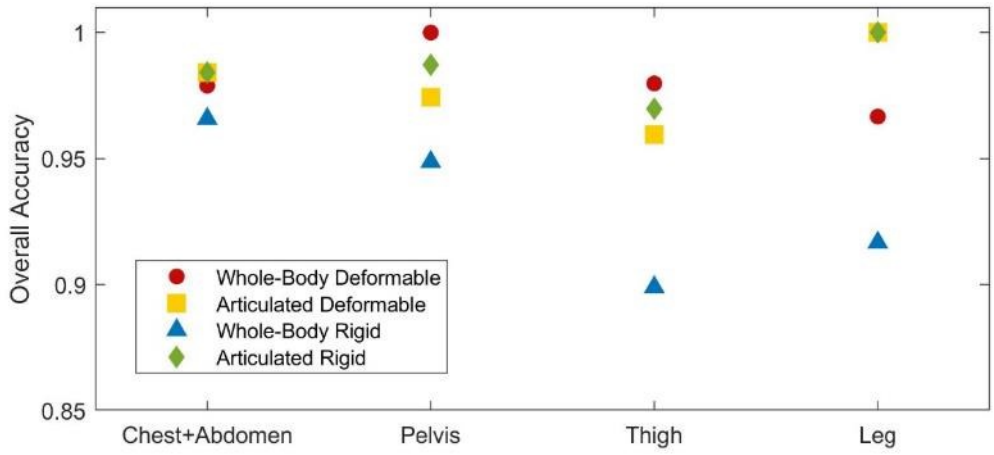


*Figure 13 – Overall accuracy, using each registration method in the four body-parts with highest number of lesions at the optimal dilations.*

#### *2.4.1.4 Lesion Burden Sensitivity Analysis*

In the complete patient population, containing all 140 scan-pairs, the median number of lesion matching decisions was 3, with minimum of 1 and maximum of 68. In the high lesion burden subset, the median number of lesion matching decisions was 19, with minimum of 11 and maximum of 68. The high lesion burden subset accounted for 345 of the 736 matching decisions (47% of the total) and for 12 of the 140 scan-pairs (8.5% of the total). Accuracy results were evaluated for the optimal dilation magnitudes (Figure 14).

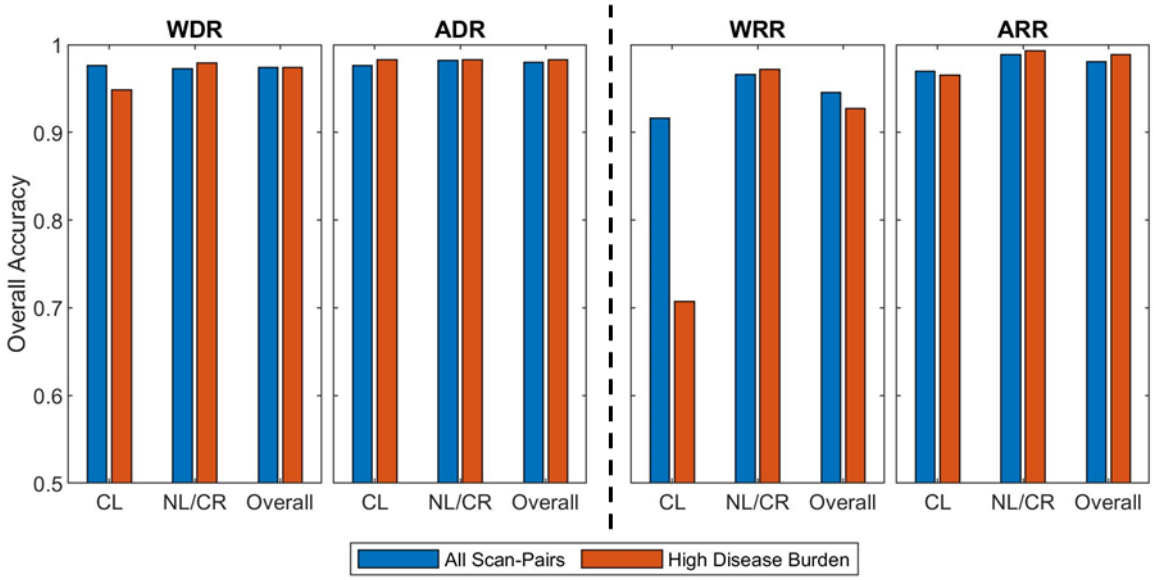


*Figure 14- Comparison between overall accuracies obtained with the whole dataset and with the subset of high disease burden scan-pairs (more than 10 matching decisions per scan-pair). Results on the left are for the 3D deformable registration methods: whole-body deformable (WDR) and articulated deformable (ADR). Results on the right are for previously explored registration methods: whole-body rigid (WRR) and articulated rigid (ARR). The optimal lesion dilation was used, namely 25 mm for ARR, ADR, and WDR; and 30 mm for WRR. In addition to overall accuracy, the analysis was separated into corresponding lesions (CL) and new or disappearing lesions (NL/DL).*

## 2.5 Discussion

In this chapter we presented a lesion matching algorithm based on 4 steps: image registration, lesion dilation, lesion clustering, and linear assignment (Figure 4). The algorithm was validated on two independent patient populations and the lesion matching accuracy obtained for the algorithm with the best performing parameters was excellent (0.98 overall accuracy). No trend in accuracy was observed across patient populations, overall accuracies were 0.98 for both. We explored the lesion matching accuracy sensitivity regarding image registration methods, lesion dilation magnitude, anatomic distribution, and lesion burden. The deformable registration techniques (WDR and ADR) showed superior registration results, when compared to rigid techniques (WRR and ARR) as shown by the landmark distance (Figure 10), which led to a significant improvement in the lesion matching accuracy. Furthermore, for WDR and ADR, residual registration error was in accordance with literature for image intensity based methods (Yin et al., 2011).

Use of lesion dilation magnitude was an important factor in achieving the excellent accuracy level. For the best performing registration approach (WDR), the use of dilation caused overall accuracy to increase by 0.08, from 0.90 with no dilation to 0.98 with the optimal dilation of 25 mm (Table 2, Figure 12). In addition, the excellent accuracy levels were minimally dependent on lesion body-part location (Figure 13) and were sustained on the analysis of high burden scan-pairs (Figure 14). Lastly, the matching criteria used by our methodology are positional and make no assumption regarding lesion shape, instead, the shape is taken into consideration by the methodology via conformal dilation (step 2) and via the use of lesion derived distance metrics for clustering (step 3).

Previous works have explored rigid registration-based lesion matching (Bosc et al., 2003; Köhler et al., 2019; Moltz et al., 2009; Shahar & Greenspan, 2005; Xu et al., 2011). Moltz et al. suggested that rigid registration suffices for lesion matching applications; however, we hypothesize that this conclusion was based on their focus on liver and lung cancer lesions. These organs are located in the chest+abdomen body-part, where we observed WRR-based matching to have its best performance (0.96 overall accuracy - Figure 13). However, for the purpose of whole-body soft tissue lesion matching, the rigid registration method (WRR) resulted in the worst overall matching accuracy; matching accuracy of corresponding lesions was of only 0.58 without dilation (Table 2). When compared to the ARR method (Yip et al., 2014), use of 3D deformable registration resulted in a remarkable lesion matching improvement, 29%, from 0.58 to 0.75 (Table 2). Deformable registration has been previously employed for the task of lesion matching (Xu et al., 2011). However, it was limited to 2D deformation of a few image slices. The use of 3D deformable registration for lesion matching purposes is one of the novelties presented in this chapter. Other novelties of our work are the use of conformal dilation for lesion matching, the development of a novel lesion clustering methodology that is specific for longitudinal monitoring of merging and splitting lesions, and the use of a non-greedy optimizer to solve a linear assignment problem in the setting of whole-body longitudinal disease matching using *in-vivo* imaging of human patients. Most importantly, the main

novelty of our work consists in combining several image analysis techniques to solve the problem of whole-body soft tissue lesion matching, which had not been done previously.

One challenge of our work was defining the reference dataset for the lesion matching decisions, which is subject to human error and bias in its nature. However, these same limitations would be encountered in manual lesion matching performed for clinical decisions. In other words, some arbitrariness is unavoidably associated with visual interpretation of medical images. We observed, however, that this limitation is likely to have little impact on the validation of the algorithm because the correspondence between the vast majority of lesion pairs was conspicuous.

The 12 incorrect matching decisions made by the algorithm for the best performer parameters (WDR with dilation of 25 mm) had three causes. First, lesion dilation caused 2 incorrect decisions where a new and a disappearing lesion were registered close enough that dilation led to incorrect superposition. Second, misregistration caused 4 incorrect decisions, all in the same scan-pair and located in the patient's legs. Largely different legs positioning resulted in a large deformation vector field that failed to accurately approximate the corresponding lesions; use of ADR for matching lesions in the legs is one way of overcoming this limitation. Finally, Incorrect clustering caused 6 incorrect decisions (Figure 15), where two or more lesions (blue) that should not cluster were superimposed to a common lesion (orange). Due to the particular geometric positioning, the threshold distance $\vec{d}$ (Figure 7D) and the angle restriction failed to prevent the incorrect clustering. This limitation could be overcome by use of an adaptive dilation magnitude parameterized by local lesion density, which was part of the methods extension presented in chapter 3. The impact of lesion matching decisions on clinical applications of the herein presented algorithm, such as treatment response prediction, will be explored in chapter 4.

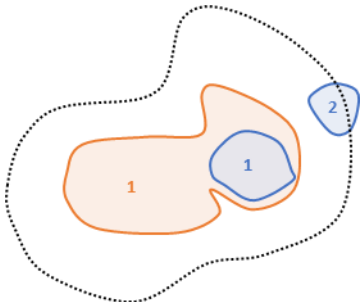


*Figure 15 - illustration of incorrect cluster decision*

While the presented algorithm was validated using PET/CT images, we hypothesize that it is generalizable to other types of medical imaging modalities. The requirements for application of the algorithm are anatomical information that is used to guide the registration step and the existence of lesion conspicuity, which in our validation dataset was obtained through PET scans. We note, however, that the implementation of the method to other imaging modalities would require sensible adaptation of hyperparameters involved in all four steps of the algorithm. These adaptations are explored in chapter 3.

## 2.6 Conclusion

In this chapter we introduced an algorithm that enables automated matching of metastatic lesions across paired images. We validated the algorithm for whole-body longitudinal PET/CT scans of metastatic cancer patients and showed it has high overall accuracy. One novelty of our algorithm is the leveraging of 3D deformable image registration, which outperformed the existing methods for every studied scenario and was a significant improvement, when compared to previously established methods. Our study suggests that whole-body lesion matching can be done automatically, possibly facilitating lesion-level disease assessment of metastatic cancer patients in clinical practice and in clinical trials, ultimately leading to enriched information for treatment response assessment and facilitating the development of novel treatment combinations.

# 3 EXPANDING THE METHODOLOGY TO LESION TRACKING

Chapter 3 describes an expansion of the lesion matching methodology to a lesion tracking framework. In lesion tracking, an arbitrary number of imaging time-points is used to track the disease evolution. In addition, other enhancements are made to the methodology, such as the ability to use multiple imaging modalities and the use of automated lesion dilation magnitudes. Material from this chapter is adapted from a manuscript titled "An automated methodology for whole-body, multimodality tracking of individual cancer lesions" submitted to Physics in Medicine and Biology in May of 2023. Furthermore, the methodology was patented through the Wisconsin Alumni Research Foundation under the title "system and method for monitoring lesion progression over multiple medical scans", number P230131US01.

## 3.1 MOTIVATION

In this chapter we aim to expand the lesion matching methodology introduced in chapter 2 in four meaningful aspects: (1) the tracking of lesions across multiple time-points, (2) across both PET/CT and PET/MRI. Additionally, improvements were made to the method to allow and facilitate its applicability to a wider set of problems, these were (3) the definition of an automated lesion dilation magnitude, conferring an anatomical location label to each lesion, and (4) locally constraining the lesion clustering step.

Methodologies for the automated lesion-level evaluation of longitudinal image series (lesion tracking) have been previously investigated. Lesion tracking is defined as establishing the correspondence between lesions across multiple time-points. Longitudinal studies of the same patient provide better information than single time-point studies because of tracking the progression of various individual lesions over time. The first developments used rigid image registration to evaluate longitudinal changes in brain anatomy captured by MR images (Freeborough et al., 1996; Hajnal et al., 1995; Lemieux et al., 1998). Subsequent developments investigated individual tracking of brain lesions identified in MR images (Bosc et al., 2003; Gerig et al., 2000;

Kikinis et al., 1999; Köhler et al., 2019; Kuckertz et al., 2021). Methodologies for tracking of cancer lesions have also been investigated for lung, liver, and lymphatic lesions using CT images (Kuckertz et al., 2022; Moltz et al., 2009; Szeskin et al., 2023; Xu et al., 2011). Yip et al. introduced a methodology to track lesions spread through the whole-body in PET/CT images, however these were constrained to bone (Yip et al., 2014). In the case of the methodology introduced in section 2, the limitation is that the evaluation is constrained to only two subsequent imaging time-points and it was only validated for soft-tissue lesions.

All previously developed methodologies use images of a single modality as input. However, metastatic cancer patients can undergo multiple imaging procedures spanning various imaging modalities (e.g., CT, MR, PET/CT, PET/MR) as part of their disease response assessment (Pirasteh et al., 2021). A methodology capable of performing individual lesion tracking across different modalities has not been developed. In this chapter, we develop a methodology capable of whole-body individual lesion tracking across an arbitrary number of imaging time-points, using multiple imaging modalities, which has the potential to improve both the understanding of disease evolution and treatment decision-making for metastatic cancer patients.

## 3.2 Methodology Description

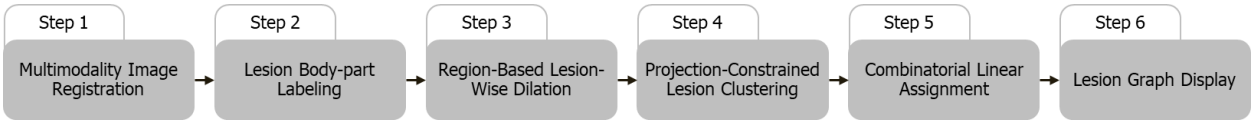


*Figure 16 – Schematic representation of the lesion tracking methodology's main steps. This schema is an expansion of the one presented in Figure 4.*

The developed lesion tracking methodology acts on an image series with associated lesion masks, i.e., it assumes that all lesions have been identified and segmented on all scans in the series and the individual lesion components are separated and numbered based on connected components analysis. The methodology performs lesion matching between each possible scan-pair within a series, and then combines the lesion matching decisions to establish lesion tracking across the entire image series. Differently from lesion tracking, lesion matching is defined as establishing the correspondence between lesions in one single scan-pair. For each

scan-pair, matching decisions are made to identify each lesion as a corresponding lesion (CL), a new lesion (NL), or a disappearing lesion (DL). The concepts of lesion tracking and matching decisions are illustrated in Figure 17. The developed methodology has six steps: (1) multimodality image registration, (2) lesion body-part labeling, (3) region-based lesion-level dilation, (4) projection-constrained lesion clustering, (5) combinatorial linear assignment, and (6) lesion graph display.

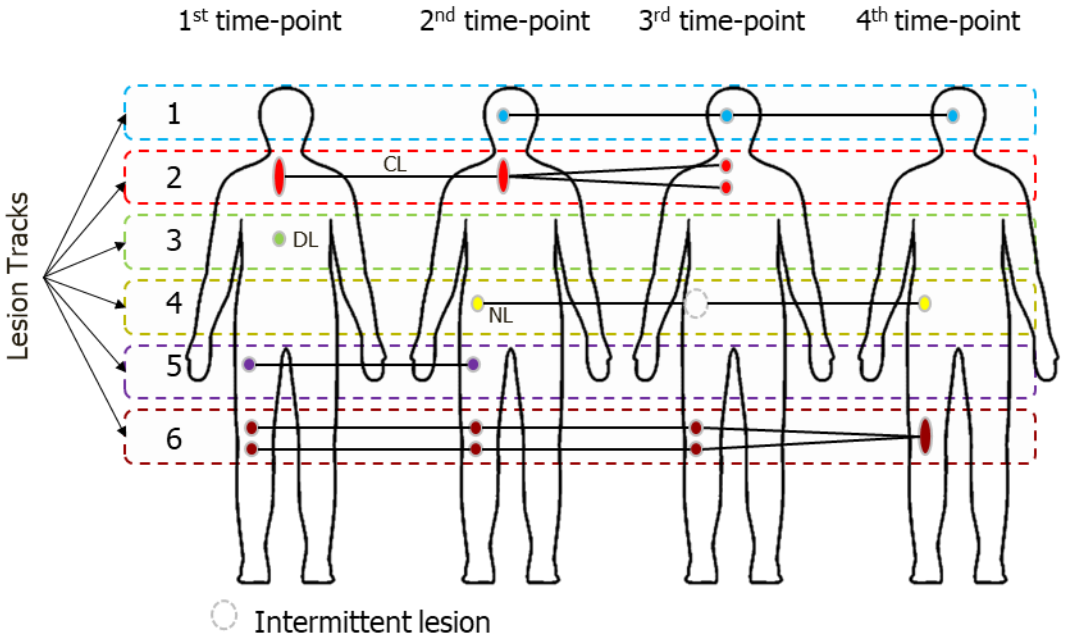


*Figure 17 – Illustrations of lesion tracks with examples of lesion merging (track 6), lesion splitting (track 2), and an intermittent lesion (track 4). The corresponding lesion (CL) matching decision is illustrated in track 2. The new lesion (NL) matching decision is illustrated in track 4. The disappearing lesion (DL) matching decision is illustrated in track 3.*

In **step 1**, all images from different time-points in a series are registered to the baseline image, resulting in $N-1$ registrations of scan-pairs, where $N$ is the number of images in a series. The intra-patient nature of the registration averts the need for registration atlases, which are commonly used for inter-patient registrations (Hellier et al., 2001). For every scan-pair, the registration mode can be either direct or indirect. Direct registration is defined as the registration between the baseline image and any other subsequent image, whereas indirect registration is defined as the registration of two subsequent images whose transformation fields were calculated to the baseline image. The registrations are performed using a whole-body rigid registration (WRR)

followed by whole-body deformable registration (WDR). The WRR performs an initial alignment of the images to prepare for the more detailed deformable step. In the WDR, a hierarchical control-grid B-spline free-form deformation (FFD) is used with a thin metal sheet bending energy regularizer (Rueckert et al., 1999). This approach has been previously validated for whole-body intra-modality image registration in section 2. The optimization metrics used are the normalized mutual information for inter-modality registration (e.g., CT to MR) (Studholme et al., 1998) and the normalized cross-correlation for intra-modality (e.g., CT to CT and MR to MR) (Avants et al., 2008).

Both metrics are optimized using a gradient descent approach (Klein et al., 2007) with the cost function

$$\hat{\mu} = argmin[C(\vec{\mu}; I_F, I_M) + \omega R(\vec{\mu})] \tag{10}$$

where $C$ denotes the optimization metric (normalized mutual information or normalized cross-correlation) as a function of the moving image ($I_M$), the fixed image ($I_F$), and the set of parameters of the image transformation ($\vec{\mu}$), which are translation and rotation parameters for WRR, and the parameters of a deformation field for the WDR (Sederberg & Parry, 1986). The function $R(\vec{\mu})$ is a penalty term (bending energy) that enforces the smoothness of the transformation (S. & Wahba, 2006). The optimal transformation field $T_n$ (for image $n$) is then applied to the coordinates of the lesion masks $L_n$ associated with each of the $N$ images in the series, generating a registered lesion mask $L_{n_T}$. The transformation field is the identity for $n = 1$ (baseline image).

$$L_{n_T} = T_n(L_n) \tag{11}$$

In **step 2** the individual lesions are labeled according to the anatomical region (body-part) where they are located. The body-parts used were head and neck, chest, pelvis, abdomen, spine, arms, and legs. Lungs were defined as a subset of the chest and liver as a subset of the abdomen. A whole-body segmentation atlas is registered to the baseline image of every series following the same registration procedure described in step 1. Every lesion is labeled according to the overlap of the lesion's volume to the body-part segmentation atlas. If

more than one body-part overlaps with a lesion volume, the lesion is labeled according to the greatest volume of overlap.

In **step 3**, the lesion contours are dilated to account for possible inaccuracies in the registration and increase the probability of lesion superposition. In section 2 a fixed dilation was used for all lesions and its magnitude was object of a sensitivity study. Here, the dilation magnitude is decided automatically and independently for each lesion and is based on the local density of lesions in the anatomical region where each lesion is situated. A morphologic conformal lesion dilation operation is applied to each lesion in an image. The lesion-specific dilation magnitude ($D_i$) is defined for each lesion as

$$D_i = \min\left(\frac{\min(d_{i,j})}{2}, D_{max}\right) \tag{12}$$

where $d_{i,j}$ is the distance between a lesion $i$ and every other lesion $j$ in the same image, and $D_{max}$ is a user defined parameter to set a maximum allowed dilation magnitude. Figure 18 illustrates the definition of $D_i$. A new dilated lesion mask $L_{n_{T,D}}$ is then defined as

$$L_{n_{T,D}} = \bigcup_i D_i\left(L_{n,i_T}\right) \tag{13}$$

where $i$ indexes each lesion in the mask.

After the dilation, the overlaps between the individual lesions ($i$ and $j$) of different time-points ($n$ and $q$) are calculated for every possible scan-pair in the series as

$$M_{n,q_{i,j}} = \left|L_{n_{T,D_i}} \cap L_{q_{T,D_j}}\right| \tag{14}$$

The matrices $M_{n,q}$ are square matrices containing the volume of overlap between the lesions of time-points $n$ and $q$. They have $\omega^2$ elements where $\omega$ is the maximal integer between the number of lesions identified in time-point $n$ ($\omega_n$) and time-point $q$ ($\omega_q$). The number of superposition matrices $M_{n,q}$ created after the whole image series evaluation is $N(N-1)/2$.

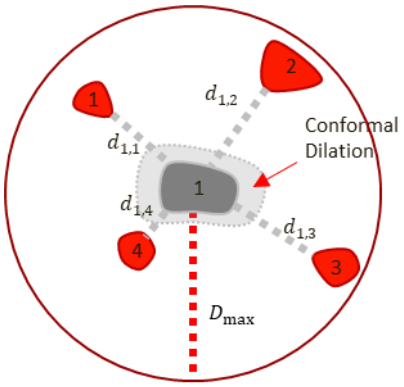


*Figure 18 – Illustration of process to determine the automated lesion dilation. All lesions belong to the same time-point.*

In **step 4,** disconnected lesion regions are clustered if they result from lesion merging or splitting. This can occur as part of the treatment response when a lesion volume shrinks and splits into many smaller lesion components or when many lesion components grow and merge as part of disease progression (Shafiei et al., 2021). A subroutine was developed to perform clustering decisions (Figure 19). The clustering decisions are made for every scan-pair by selecting the lesion pairs that have a common intersecting lesion and using geometrical considerations to decide if these pairs consist of separate lesions or if they are components of the same lesion that have merged or split. In short, lesions are clustered if a characteristic distance between them ($u$) is smaller than the longest chord of the common intersecting lesion ($d$). The lesion clustering subroutine consists of the sub-steps in Table 4. The main difference between the clustering approaches presented in sections 2 and 3 is the projection constraining aspect. This aspect is vital for correctly clustering lesions with markedly different volumes. Without the constraining step, the chord $d$ is likely to be overestimated when the intersecting lesion is much bigger than the lesion being considered for clustering.

*Table 4 – Description of the lesion clustering subroutine.*

| Sub-step | Description | Rationale |
|---|---|---|
| **4.1** | Determines a connecting orientation $\hat{t}$ between the lesion pair | The distances $u$ and $d$ will be measured along this orientation |
| **4.2** | Defines a characteristic separation distance $u$ between both lesions in the pair, where $u$ is aligned with the orientation $\hat{t}$ | The distance $u$ characterizes the separation of the lesion pair |
| **4.3** | Projects the lesion pair contours along $\hat{t}$ to constrain the common intersecting lesion contour $C$ | The measurement of $d$ needs to be locally constrained (see $d^*$ in Figure 19D) |
| **4.4** | Determines the length $d$ of the longest chord within the constrained common intersecting lesion contour ($C_C$), where $d$ is aligned with the orientation $\hat{t}$ | The distance $d$ defines a limit separation assuming the lesion pair originated from lesion merging (or splitting) |
| **4.5** | Compares $d$ and $u$ to make the clustering decision (positive for cluster if $d > u$) | Makes the final clustering decision |

In step 4.1, the orientation $\hat{t}$ is defined as the weighted mean of all the vectors connecting each voxel of both lesions in the pair

$$\hat{t} = \frac{1}{V}\sum_{i=1}^{N} \vec{t}_i \cdot w_i \tag{15}$$

where $\vec{t_i}$ is a connecting vector, $V$ is the total number of connecting vectors, and the weights $w_i$ are defined as the inverse L2 norm of each vector

$$w_i = \frac{1}{|\vec{t}_i|} \tag{16}$$

In step 4.2, the distance $u$ is determined as the 95$^{th}$ percentile of the distribution of the norms of all vectors $\vec{t_i}$ that satisfy the angular constraint

$$u = P_{95}(\vec{t}_i), \text{ subject to } \quad \arccos\left(\frac{\vec{t}_i \cdot \hat{t}}{|\vec{t}_i| \cdot |\hat{t}|}\right) \leq 5°. \tag{17}$$

this imposes a soft constraint that the vectors being considered to determine $u$ are aligned with $\hat{t}$. The 5 degrees value is a hyperparameter of the methodology that can be adapted according to each application. The 95$^{th}$ percentile was chosen instead of a maximal operation because it is more robust against outliers in the distribution (Huff et al., 2021).

In step 4.3, a constrained version (denoted by $C_c$) of the common intersecting lesion contour ($C$) is determined by projecting the contours ($L_i$ and $L_j$) of both lesions in the pair in the orientation $\hat{t}$, generating the projected contours ($L_{i,p}$ and $L_{j,p}$). $C_c$ is then defined as

$$C_c = L_{i,p} \cap L_{j,p} \cap C \tag{18}$$

$C$ needs to be constrained to account for lesions of significantly different dimensions (as demonstrated in Figure 19D).

In step 4.4, $\vec{d}$ is determined as the longest of the chords $\vec{d}_i$ within $C_c$ subject to an angular constraint,

$$d = max(|\vec{d}_i|), \text{ subject to } \arccos\left(\frac{\vec{d}_i \cdot \hat{t}}{|\vec{d}_i| \cdot |\hat{t}|}\right) \leq 5°. \tag{19}$$

this enforces again that the chord $\vec{d}$ is oriented in a similar orientation as $\hat{t}$, hence the distances $d$ and $u$ are both measured along similar orientations.

Finally, in step 4.5, $u$ and $d$ are compared, and if $u < d$ then the lesion pair is considered as a cluster, rather than as two disconnected lesions. Figure 19 illustrates the clustering sub-routine. After step 4, the superposition matrices $M_{n,q}$ for the whole series are reevaluated based on the altered lesion masks, generating modified matrices $M'_{n,q}$.

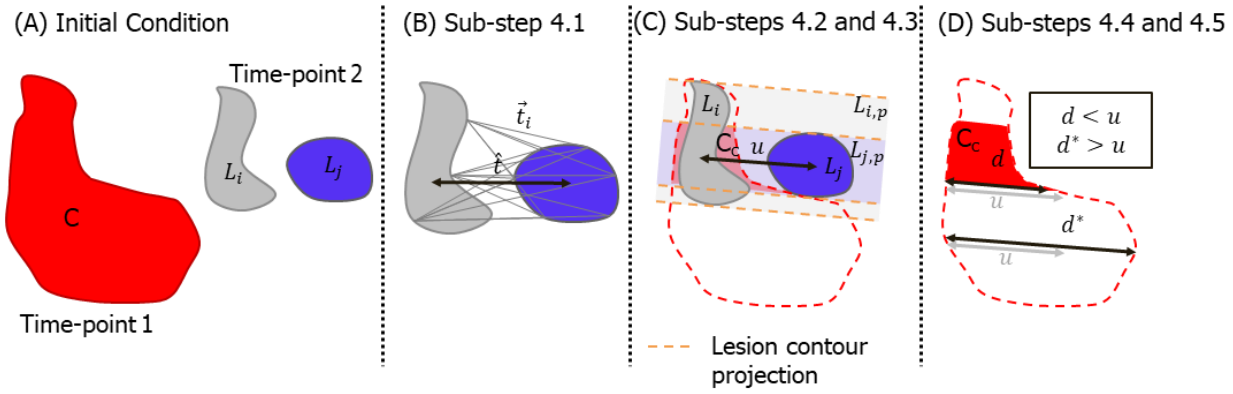


*Figure 19 – Illustration of the clustering subroutine. (A) shows the lesions in the initial condition, where a lesion $C$ has shrunk to a much smaller lesion $L_i$ in the time-point 2 and a new lesion $L_j$ appeared. In (B) the direction $\hat{t}$ is determined. In (C), the characteristic distance $u$ is determined and $L_i$ and $L_j$ are projected to create the constrained contour $C_c$. In (D) the longest chord distance $d$ is determined and compared to $u$. The lesions $L_i$ and $L_j$ are not clustered since $d < u$. Importantly, lesions $L_i$ and $L_j$ would be wrongfully clustered if the lesion projection was not considered since $d^* > u$.*

In **step 5**, lesion matching decisions are made by evaluating every possible scan-pair combination in the series and establishing lesion correspondence. The combinatorial linear assignment process defines a cost matrix $\mathrm{K}$ for each matrix $M'_{n,q}$ as $\mathrm{K}_{n,q_{i,j}} = 1/M'_{n,q_{i,j}}$. Lesion matching is reduced then to a linear assignment problem of finding the optimal permutation matrices $A_{n,q}$ that globally maximize the overlap between the lesions in the scan-pair (Jaqaman et al., 2008)

$$A_{n,q} = \underset{A_{n,q}}{\operatorname{argmin}} \left( \sum_{i=1}^{\omega} \sum_{j=1}^{\omega} A_{n,q_{i,j}} \mathrm{K}_{n,q_{i,j}} \right) \tag{20}$$

within the constraints to the permutation matrix $A$

$$\textstyle\sum_{i=1}^{\omega} A_{n,q_{i,j}} = 0 \text{ or } 1 \text{ and } \sum_{j=1}^{\omega} A_{n,q_{i,j}} = 0 \text{ or } 1 \tag{21}$$

where the sum of the matrix elements can be zero if there are no matches for the lesions represented by those rows and columns. The Munkres assignment algorithm (Munkres, 1957) is used to find each optimal matrix $A_{n,q}$ via a non-greedy minimization of the cost matrices $\mathrm{K}_{n,q}$.

Finally, in **step 6**, the lesion tracking information of the image series is summarized as a lesion graph (Kuckertz et al., 2021; Szeskin et al., 2023; Yan et al., 2018). This structure has one layer per image in the series and each node represents one lesion in each time-point, if many lesion components are clustered, each component is still represented by an individual node in the lesion graph. The edges represent the temporal correspondence between the lesions, i.e., the lesion matches. The lesion graph corresponding to the illustration of Figure 17 is shown in Figure 20. A lesion track is defined by a connected subset of the graph that is disconnected from every other subset.

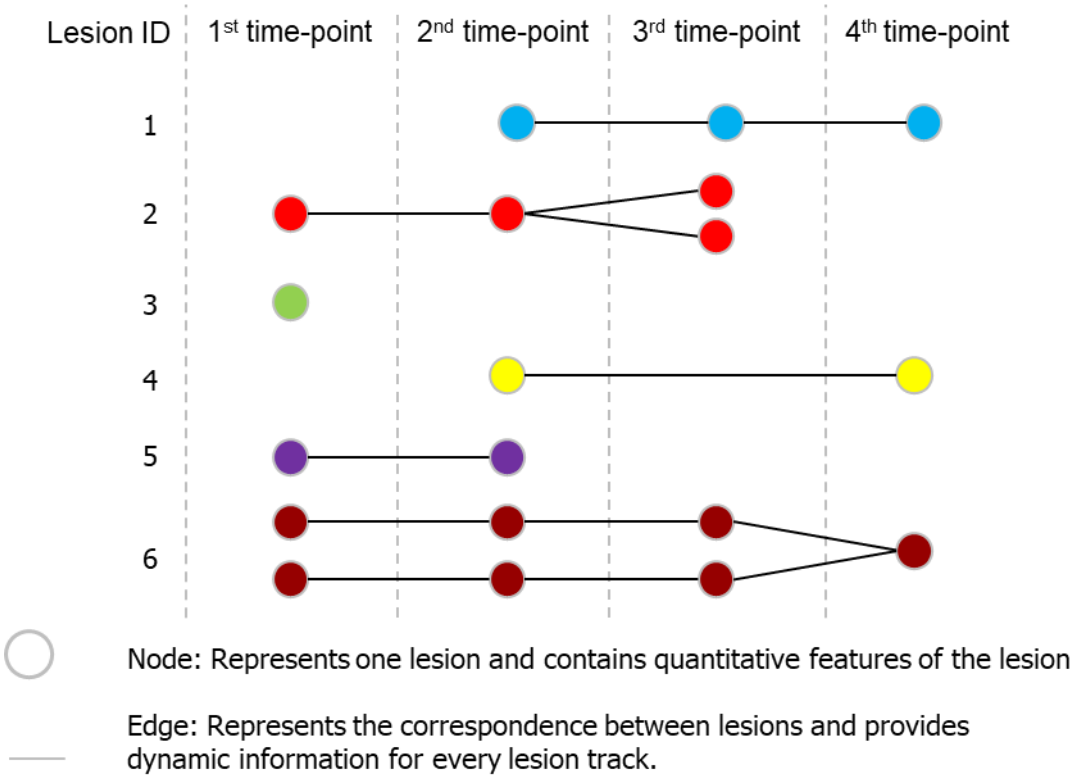


*Figure 20 – Lesion graph of the hypothetical patient exhibited in Figure 17. Each lesion track is presented in a different color.*

## 3.3 Methodology Benchmarking and Sensitivity Analyses

The methodology was benchmarked on a population of mNET patients. A mNET patient population was used for methodology benchmarking in this chapter to expand on the cancer types validated in chapter 2, since NET was not included there. More importantly, mNET patients' treatment response assessment is of interest

given the significantly different responses these patients present to PRRT (Werner et al., 2017). The patient response to PRRT and how comprehensive, lesion-level, longitudinal analysis impacts the predictability of treatment response will be studied in detail in chapter 4. Furthermore, some NET characteristics pose relevant challenges to the proposed methodology, and thus serves as an appropriate test set for benchmarking. First, NET diagnosis and staging relies on multiple anatomical and molecular imaging modalities (Pirasteh et al., 2021). Second, NET patients undergo numerous imaging procedures due to the tumors' slow growing nature and patients' relatively long overall survival (5 year survival rate of 77%) (Figiel et al., 2020). Third, the patients present lesions both in the soft-tissue and in bones. The patient cohort was retrospectively collected from the University of Wisconsin Hospitals and Clinics Picture Archiving and Communication System (UWHC-PACS). All the ethical guidelines and requirements of the institutional review board were followed. The inclusion criteria were metastatic NET patients that had received both PET/CT and PET/MR whole-body scans as part of their care in UWHC. PET/CT and PET/MR were used because they have whole-body fields-of-view (FOVs) and because the PET image provides great lesion conspicuity. In this chapter we have constrained the patient population to those that received both PET/CT and PET/MR because one of our goals was to benchmark the lesion tracking methodology using various imaging modalities. The search returned 6 patients that were scanned a total of 15 times, resulting in 7 PET/MR and 8 PET/CT scans. The PET/CT images were acquired on the scanners GE Discovery 710, GE Discovery MI, and Siemens Biograph mCT. The PET/MR images were acquired on a GE Signa Scanner. The PET tracer for all images was $^{68}$Ga-DOTATATE. The median injected dose was 162 (range: 118 to 467) MBq, and the median time interval from radiopharmaceutical injection to image acquisition was 83 minutes (range: 53 to 101). More details about the image acquisitions are presented in Table 5. The images had varying voxel-sizes and all images were resampled to 2x2x2 mm.

*Table 5 – Patient population and imaging information*

| Patient ID | Time-Point | Inj. dose (MBq) | Scanner | Type | Δtime (weeks) | # of Lesions |
|---|---|---|---|---|---|---|
| 1 | 1 | 175 | GE Discovery 710 | PET/CT | 0 | 106 |
| | 2 | 189 | GE Signa | PET/MR | 51 | 90 |
| | 3 | 168 | GE Discovery MI | PET/CT | 73 | 128 |
| | 4 | 150 | GE Signa | PET/MR | 101 | 83 |
| 2 | 1* | 467 | GE Discovery LS | PET/CT | 0 | 16 |
| | 2* | 198 | Siemens Biograph mCT | PET/CT | 75 | 44 |
| | 3 | 155 | GE Signa | PET/MR | 183 | 26 |
| 3 | 1 | 145 | GE Signa | PET/MR | 0 | 34 |
| | 2 | 202 | GE Discovery MI | PET/CT | 39 | 70 |
| 4 | 1 | 150 | GE Discovery MI | PET/CT | 0 | 83 |
| | 2 | 118 | GE Signa | PET/MR | 37 | 42 |
| 5 | 1 | 137 | GE Discovery MI | PET/CT | 0 | 7 |
| | 2 | 194 | GE Signa | PET/MR | 64 | 14 |
| 6 | 1 | 162 | GE Discovery MI | PET/CT | 0 | 27 |
| | 2 | 144 | GE Signa | PET/MR | 91 | 20 |

*Image acquired outside UWHC and imported to UWHC-PACS

The cancer lesions were identified and contoured manually. This process was performed by an image scientist and supervised by a nuclear medicine physician. Segmentation was performed on the fused PET/CT images of all time-points by observing and manually segmenting regions of conspicuous elevated tracer uptake in relation to the local tracer background. Volumes of naturally occurring elevated uptake were not included in the segmentations. The anatomical information contained in the CT and MR scans was used to compute the image registration transformation fields, which were then applied to the binary lesion masks. The registration step was implemented using elastix (Klein et al., 2010; Shamonin et al., 2014) with the SimpleElastix python wrapper (Marstal et al., 2016) and using a variation of the gain factor variables published by Akbarzadeh et al. (Akbarzadeh et al., 2013). The methodology's steps 2 through 6 were implemented in MATLAB version R2022b (The Mathworks Inc.).

### 3.3.1 Gain Factor Optimization

In the gradient descent framework used to calculate the deformable registration vector field, the parameter set ($\vec{\mu}$) is iteratively optimized according to the equation

$$\vec{\mu}_{k+1} = \vec{\mu}_k - a_k \left|\frac{\partial C}{\partial \vec{\mu}}\right|_{\vec{\mu}_k} \tag{22}$$

where $k$ is the iteration index and $a_k$ is a decaying gain factor defined by

$$a_k = \frac{a}{(A+k)^\alpha} \tag{23}$$

with variables $a > 0$, $A \geq 1$ and $0 < \alpha < 1$ (Klein et al., 2007).

The gain factor variables defined by Akbarzadeh et al. were used in our work as references, around which an optimization was employed to tailor the registration to our needs (Akbarzadeh et al., 2013). In their work, gross anatomical mismatches, such as different head positions, were not accounted for. These can differ considerably for PET/CT and PET/MR scanners. Hence the motivation for further optimization of the gain factor variables.

The gain factor variables were optimized using a random sampling framework. The deformable registration hierarchical control-grid had 5 layers with different grid spacing and gaussian blurring intensities (image resolution). Each layer had its own gain factor variables that were optimized independently. In the optimization, sets of gain factor variables were sampled from intervals defined around the variables published by Akbarzadeh et al. To avoid redundancy, a sampled set was rejected if any variable was too close to any other variable previously selected, where closeness was determined by a predefined rejection distance. For each layer, the final set of variables selected was the one that minimized the image similarity cost function (normalized mutual information). A list of the reference grid parameters, reference gain factor variables, sampling intervals, and rejection distances is shown in Table 6.

*Table 6 – Reference parameters, intervals, and rejection distances used in the random sampling optimization of gain factors (equation 23).*

| | Description | Layer 0 | Layer 1 | Layer 2 | Layer 3 | Layer 4 |
|---|---|---|---|---|---|---|
| Grid parameters | Spacing (mm) | 120 | 60 | 60 | 30 | 30 |
| | $\sigma$(mm) | 2 | 1.5 | 1 | 0.5 | 0.5 |
| Reference gain factors (Akbarzadeh et al) | $a$ | 2000 | 7000 | 20000 | 20000 | 30000 |
| | $A$ | 50 | 50 | 50 | 50 | 50 |
| | $\alpha$ | 0.602 | 0.602 | 0.602 | 0.602 | 0.602 |
| Sampling Intervals | $a$ | [500, 4000] | [1750, 14000] | [5000, 40000] | [5000, 40000] | [7500, 60000] |
| | $A$ | [100, 400] | [100, 400] | [100, 400] | [100, 400] | [100, 400] |
| | $\alpha$ | [0.1, 0.6] | [0.1, 0.6] | [0.1, 0.6] | [0.1, 0.6] | [0.1, 0.6] |
| Rejection distances | $a$ | 100 | 350 | 1000 | 1000 | 1500 |
| | $A$ | 5 | 5 | 5 | 5 | 5 |
| | $\alpha$ | 0.01 | 0.01 | 0.01 | 0.01 | 0.01 |

### 3.3.2 Image Registration Accuracy and Sensitivity Analyses

The distances between the centroids of landmark lesions were used to evaluate the accuracy of image registration. A lesion was used as landmark when there was no more than 50% volume variation of that lesion within the scan-pair being registered. This volume constraint was introduced to limit the biasing in the expected value of the centroid difference by lesions significantly growing or shrinking. Three registration accuracy sensitivity analyses were performed. The first compared the WRR and the WDR accuracies. The second compared direct vs. indirect registration modes. The third sensitivity analysis was done to compare the accuracies of intra-modality (CT-CT and MR-MR) against inter-modality registration (CT-MR).

### 3.3.3 Performance of Lesion Matching and Tracking

Reference lesion matches were manually prepared by an imaging scientist (VSF) and served to evaluate the methodology's lesion matching and lesion tracking performances. Matching accuracy was evaluated as the number of correct matching decisions divided by the total number of reference matching decisions. Accuracy was separately determined for corresponding lesions (CL) and for new and disappearing lesions (NL/DL). To evaluate the lesion tracking accuracy, the number of automated lesion tracks identified by the methodology

with an identical reference track was determined. An identical reference track means that the automated and the reference tracks contain the same lesions and lesion matches. The identical track ratio was determined by dividing the number of identical tracks by the total number of reference tracks.

Another way to evaluate the matching performance is the precision/recall (or PPV/sensitivity) approach. In this approach, every possible lesion pair was evaluated. If they were included in the same lesion track, the decision was a TP or a FP (depending on automated-reference agreement). If they were not included in the same track, the decision was a TN or a FN. The F-score was calculated from the sensitivity and the PPV as

$$F = \frac{2}{\frac{1}{sens} + \frac{1}{PPV}} \tag{24}$$

### 3.3.4 Sensitivity Analyses

#### *3.3.4.1 Lesion Dilation Sensitivity Analysis*

The lesion matching and lesion tracking accuracies were evaluated by employing various fixed magnitudes of lesion dilation (0 to 18 mm in 3 mm steps) and using the automated dilation (as described in section 3.2, step 3) to determine an optimal dilation strategy.

#### *3.3.4.2 Anatomic Distribution Sensitivity Analysis*

A reference set of body-part labels was manually established by an imaging scientist (author VS) and the lesion body-part labeling accuracy was determined as the number of correct automated labels divided by the total number of lesions. The lesion body-part labeling, the lesion matching, and the identical track ratio accuracies were evaluated in relation to the body-part in which the lesions were located. A body-part labeling confusion matrix was prepared (Stehman, 1997).

#### *3.3.4.3 Image Modality Sensitivity Analysis*

The impact on lesion matching accuracy was evaluated with respect to the image modalities involved in the registration step of the methodology. This was done by inter-comparing the lesion matching accuracy of CT-MR, CT-CT, and MR-MR registrations. The identical track ratio could not be used in the image modality sensitivity analysis. To measure the identical track ratio, every scan in a patient's series is necessary. Therefore, modalities cannot be separated.

#### *3.3.4.4 Registration Mode Sensitivity Analysis*

The lesion matching accuracy was evaluated for the direct and the indirect registration modes. The identical track ratio could not be used in the registration mode sensitivity analysis. To measure the identical track ratio, every scan in a patient's series is necessary. Therefore, registration modes cannot be separated.

#### *3.3.4.5 Track Size Sensitivity Analysis*

The identical track ratio was evaluated as a function of the number of layers in the track and of the number of lesion contours included in the track (number of nodes).

#### *3.3.4.6 Failure Modes*

The reason for every false lesion matching decision made by the automated lesion tracking methodology was investigated and a qualitative analysis was performed to group the errors into similar failure modes.

### 3.3.5 Statistical Analysis

Statistical analysis was performed across the studies described in Sections 3.3.1 to 0. For the evaluation of registration accuracy, the direct and indirect registration landmark distributions were compared using the Wilcoxon signed-rank test. For the testing of different imaging modalities, a Kruskal-Wallis' test was used to evaluate if all distributions could be drawn from the same underlying distribution, followed by a Wilcoxon rank test for difference in the distributions, if necessary. The significance level considered was $p < .05$. The Bonferroni method was used to correct for multiple comparisons. The patient-wise standard deviation was

calculated for matching accuracies and identical track ratios. The patient-wise Pearson correlation coefficient between the overall accuracy and the identical track ratio was calculated to evaluate the degree of dependence between lesion matching and lesion tracking accuracies. The standard deviations of the accuracies and identical track ratios were calculated across patients and across lesion body-parts as a metric of the methodology's robustness. In the intra- vs. inter-modality registration accuracy analysis, the image modality sensitivity analysis, and the registration mode sensitivity analysis, only images from the patients who had more than two images in their series were used (patients 1 and 2) to avoid category imbalance.

## 3.4 RESULTS

There were 790 unique lesion contours in the patient population. After visual analysis, lesions were grouped into 396 reference lesion tracks consisting of 1,183 reference matching decisions. The results per patient are shown in Table 7.

*Table 7 – Results from the reference lesion matching analysis, used to benchmark the methodology.*

| Patient ID | Images in Series | Individual Lesions | Ref. Tracks | Ref. Matching Decisions |
|---|---|---|---|---|
| 1 | 4 | 407 | 145 | 812 |
| 2 | 3 | 86 | 54 | 141 |
| 3 | 2 | 104 | 75 | 87 |
| 4 | 2 | 125 | 88 | 98 |
| 5 | 2 | 21 | 8 | 15 |
| 6 | 2 | 47 | 26 | 30 |
| **Total** | **15** | **790** | **396** | **1,183** |

### 3.4.1 Gain Factor Optimization

Table 8 shows the parameter optimization results for the deformable registration process. The optimization procedure was performed for a single pair of images (CT-MR). However, a noticeable qualitative improvement in registration was observed for all patients and scan-pairs.

*Table 8 – Gain factor variables results from the random sampling optimization.*

| Parameters | Layer 0 | Layer 1 | Layer 2 | Layer 3 | Layer 4 |
|---|---|---|---|---|---|
| $a$ | 3500 | 14000 | 20000 | 20000 | 31000 |
| $A$ | 290 | 150 | 150 | 300 | 10 |
| $\alpha$ | 0.004 | 0.017 | 0.073 | 0.007 | 0.403 |

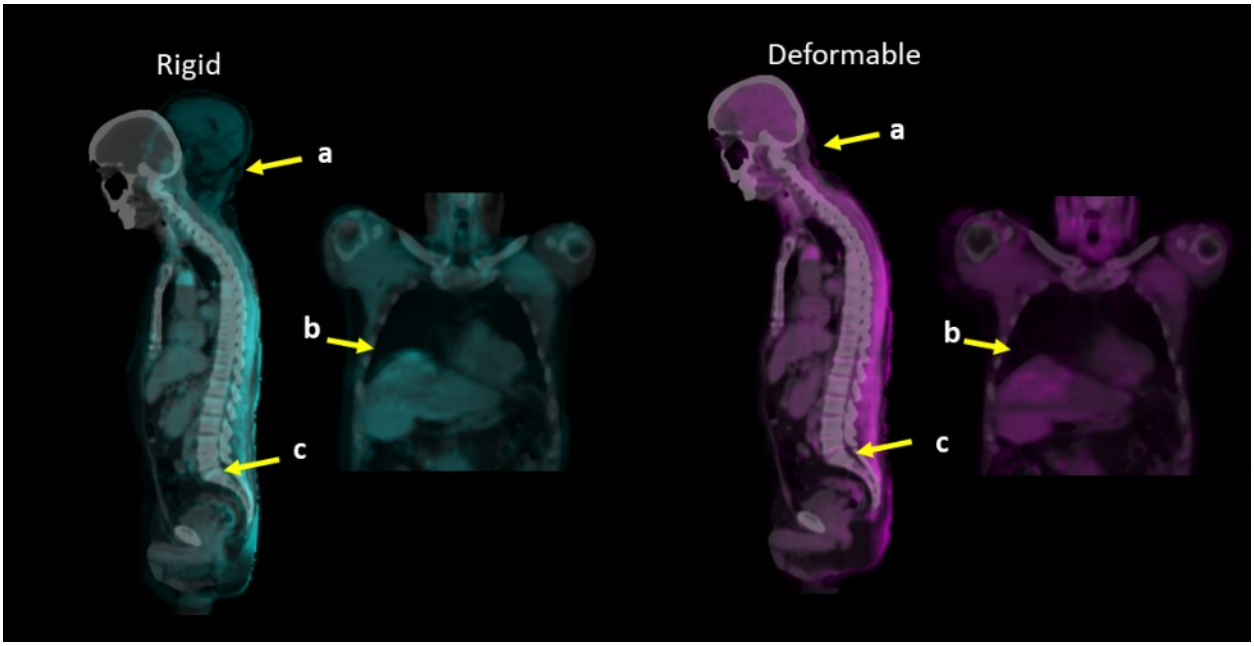


*Figure 21 – sagittal and coronal slices of qualitative demonstration of anatomical alignment improvement from rigid (RHS) to deformable registration (LHS). Arrows indicate (a) improvement in head position registration, (b) better liver superposition, and (c) better spine alignment.*

#### *3.4.1.1 Image Registration Accuracy and Sensitivity Analyses*

Out of the 790 lesions in the dataset, 291 (37%) fulfilled the volume variation constraint to be landmark lesions. The median landmark centroid distance (Figure 22A) using WRR was 4.2 mm and using WDR it was 2.6 mm. The registration accuracy for WDR was significantly higher than for WRR ($p < 0.001$). The two most prominent outliers seen on the WDR distribution (25.7 and 26.7 mm) come from lesions whose diameters are considerably larger than $25$ mm (Figure 23), therefore these registrations can still be considered acceptable.

The results for the comparison between direct and indirect WDR are shown in Figure 22B. Direct and indirect registrations had non-distinguishable performances (p = 0.1).

Finally, the intra- and inter-modality distributions of centroid distances (Figure 22C) were significantly different. The Kruskal-Wallis p-value for the Chi-Square test was $p \ll 0.001$, the pairwise tests were done using a Bonferroni-corrected significance level of 0.02. Only patients who received all three possible modality combinations (patients 1 and 2) were included in the inter-modality analysis, consequently 265 of 291 lesion pairs were used. The median centroid distance for CT-MR, MR-MR, and CT-CT registrations were 2.9, 2.4, and 1.7 mm, respectively. Inter-modality (CT-MR) registration performance was significantly worse than intra-modality performance (CT-CT and MR-MR). Both intra-modality registration performances (CT-CT and MR-MR) were indistinguishable.

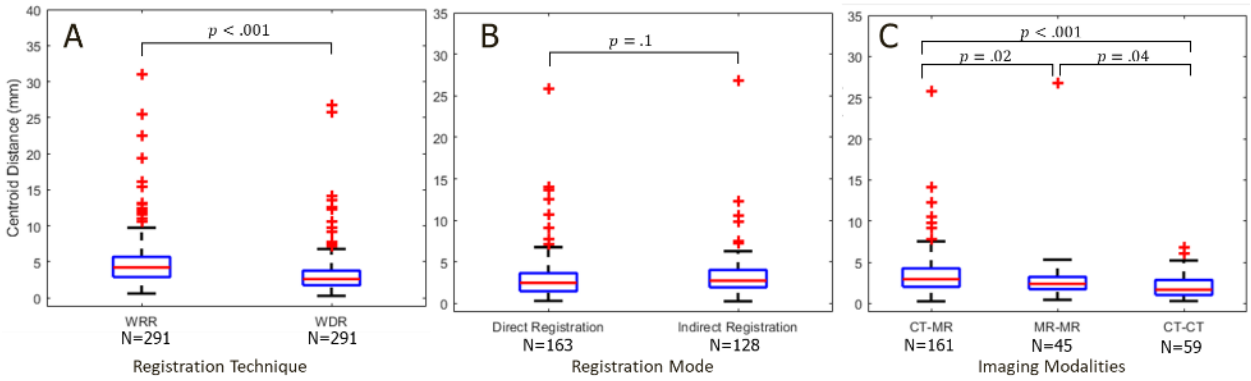


*Figure 22 – Registration performance sensitivity analysis with respect to (A) registration technique, comparing whole-body rigid registration (WRR) and whole-body deformable registration (WDR) techniques. (B) by direct vs. indirect registration. (C) by modalities involved in the registration.*

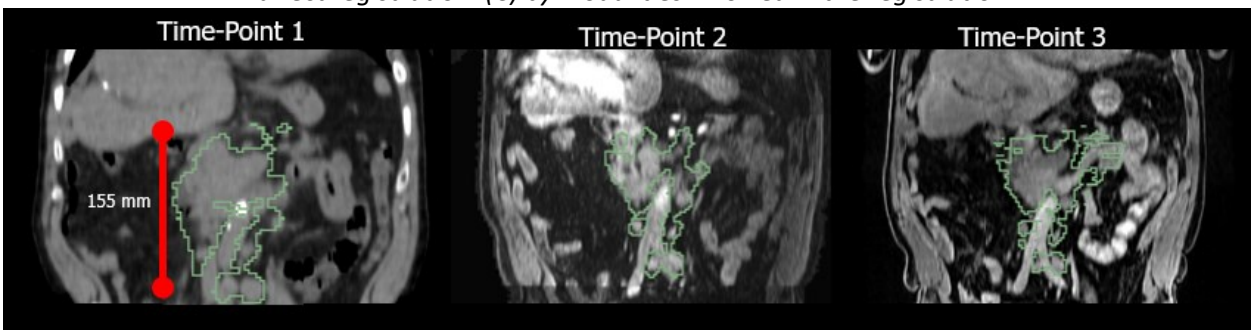


*Figure 23 – Coronal slice of abdominal region displaying the lesion involved in causing centroid distance outliers observed in Figure 22A ($\geq 25.7\ mm$). The PET signal was omitted to improve visibility of the anatomic structures.*

### *3.4.1.2 Performance of Lesion Matching and Tracking*

WDR was used for all lesion matching and tracking performed in this chapter's analyses since it showed superior registration accuracy (than WRR) and was previously demonstrated to yield optimal lesion matching

results (Santoro-Fernandes et al., 2021) (also shown in Section 2.4). Without the use of lesion dilation (step 3), the identical track ratio was 85%, the overall matching accuracy was 0.86, CL matching accuracy was 0.75 and NL/DL matching accuracy was 0.99. In terms of precision/recall, the lesion matching method achieved 0.89 recall (sensitivity) and 0.90 precision (PPV) leading to an F-score of 0.90, when no lesion dilation was used.

### 3.4.2 Sensitivity Analyses

#### *3.4.2.1 Lesion Dilation Sensitivity Analysis*

The automated lesion matching accuracy and identical track ratio for each fixed dilation magnitude and for the automated dilation approach are shown in Figure 24. In general, the automated dilation magnitude surpassed any of the fixed dilation magnitude results, leading to 356 out of 396 identical tracks (90% identical track ratio) and an overall accuracy of 0.91. The CL accuracy for 9 and 12 mm dilations was better than for the automated dilation case, however the slightly better CL accuracy was offset in the overall accuracy by the inferior NL/DL accuracy. Automated dilation led to CL accuracy of 0.83 and NL/DL accuracy of 1.

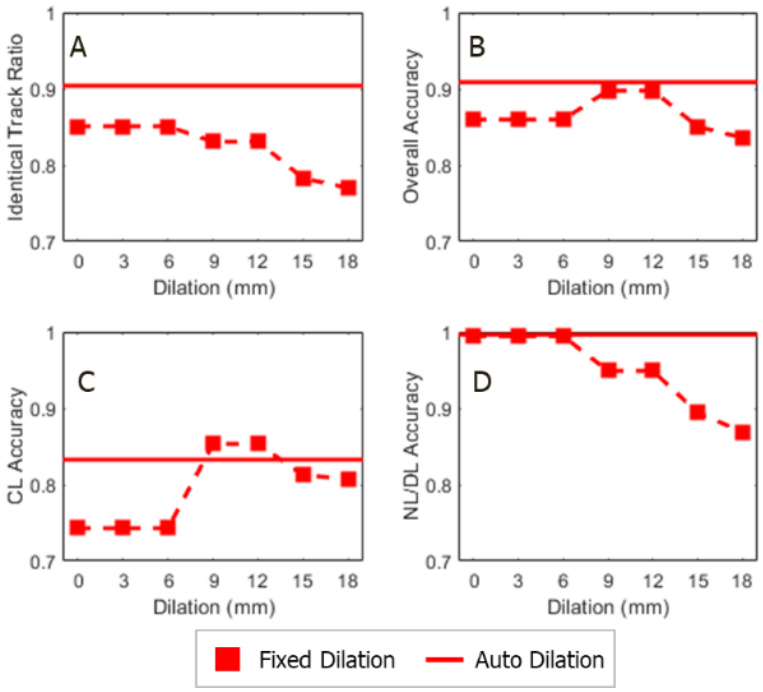


*Figure 24 – Results for (A) Identical track ratio, (B) overall lesion matching accuracy, (C) corresponding lesions (CL) accuracy, and (D) new or disapearing lesion (NL/DL) accuracy.*

Table 9 shows the performance metrics per patient in the automated dilation case. The correlation between the overall accuracy (indicative of lesion matching performance) and the identical track ratio (indicative of lesion tracking performance) was $\rho = 0.3$.

*Table 9 – Lesion matching and lesion tracking accuracies shown by patient. Results using automated dilation.*

| Patient ID | CL Acc. | NL/DL Acc | Overall Acc. | Identical Tracks Ratio |
|---|---|---|---|---|
| 1 | 0.84 | 0.99 | 0.90 | 86% |
| 2 | 0.66 | 1.00 | 0.78 | 89% |
| 3 | 0.90 | 1.00 | 0.93 | 96% |
| 4 | 0.84 | 1.00 | 0.89 | 93% |
| 5 | 0.85 | 1.00 | 0.92 | 75% |
| 6 | 1.00 | 1.00 | 1.00 | 100% |
| **Total** | 0.83 | 1.000 | 0.91 | 90% |
| **Std. dev.** | 0.11 | 0.004 | 0.07 | 9% |

Figure 25 shows the lesion matching results as a function of the dilation magnitude in terms of precision (PPV) and recall (sensitivity).

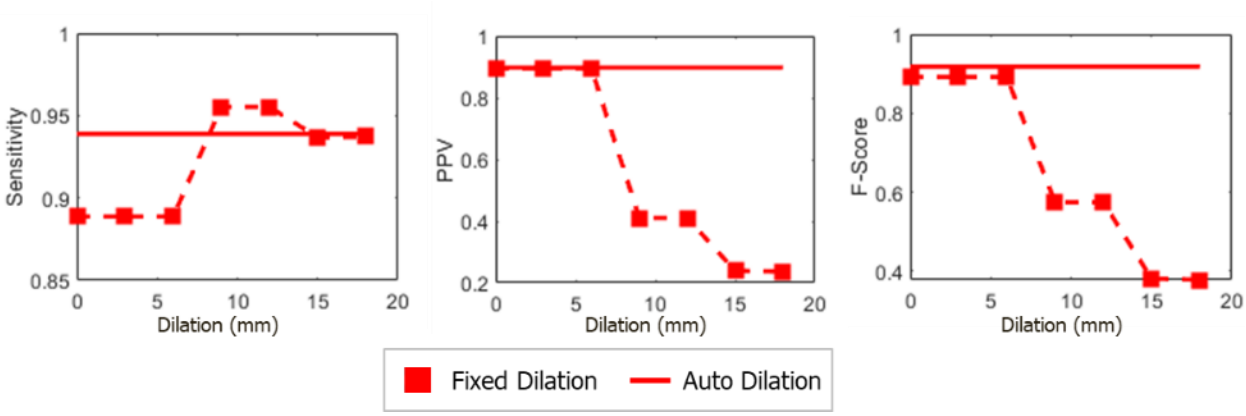


*Figure 25 – Lesion matching results in terms of sensitivity, PPV, and F-Score.*

### *3.4.2.2 Anatomic Distribution Sensitivity Analysis*

The 790 lesions were distributed from head to thighs, with greater prevalence in the abdominal area. The automated lesion body-part determination step had a sensitivity of 0.86, correctly labeling 678 lesions out of the 790, Figure 26 shows the confusion matrix for lesion body-part labeling. Table 10 shows the number of lesions and lesion tracking performance metrics per body-part.

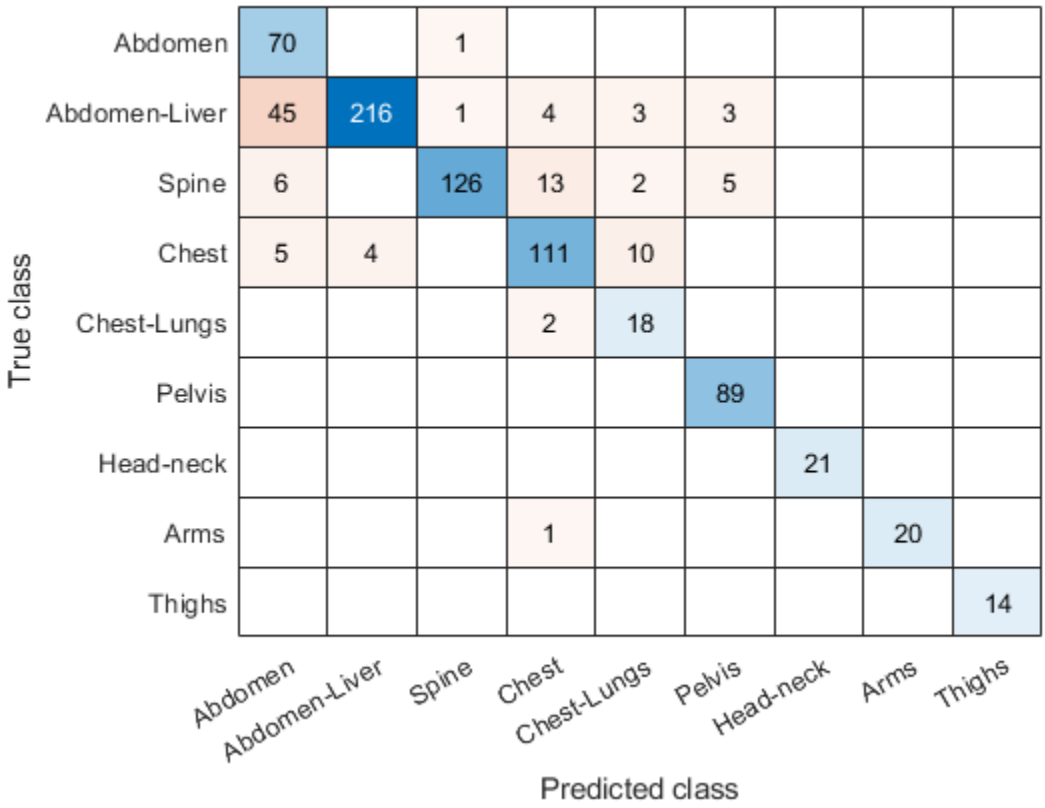


*Figure 26 – Confusion matrix for lesion body-part labeling. "Abdomen-Liver" is a sub-category of the "Abdomen" category, and "Chest-Lungs" is a sub-category of the "Chest" category.*

*Table 10 - Results of lesion body-part sensitivity analysis, using automated dilation.*

| Body-part | # of lesions | Labeling Sensitivity | CL Acc. | NL/DL Acc. | Overall Acc. | Identical Tracks Ratio |
|---|---|---|---|---|---|---|
| Abdomen | 343 (43%) | 0.83 | 0.81 | 1.00 | 0.91 | 91% |
| *Liver* | *272 (34%)* | *0.79* | *0.82* | *1.00* | *0.92* | *93%* |
| Spine | 152 (19%) | 0.83 | 0.82 | 1.00 | 0.88 | 89% |
| Chest | 150 (19%) | 0.81 | 0.90 | 1.00 | 0.94 | 97% |
| *Lungs* | *20 (3%)* | *0.90* | *0.94* | *1.00* | *0.97* | *100%* |
| Pelvis | 89 (11%) | 1.00 | 0.84 | 1.00 | 0.91 | 100% |
| Head-neck | 21 (3%) | 1.00 | 0.77 | 0.94 | 0.89 | 82% |
| Arms | 21 (3%) | 0.95 | 0.73 | 1.00 | 0.90 | 78% |
| Thighs | 14 (2%) | 1.00 | 0.91 | 1.00 | 0.96 | 88% |
| **Total** | 790 | 0.86 | 0.83 | 1.00 | 0.91 | 90% |
| **Std. dev.** | | 0.09 | 0.07 | 0.02 | 0.03 | 8% |

#### *3.4.2.3 Image Modality Sensitivity Analysis*

When evaluating the dependence of lesion tracking performance with imaging modality involved in the registration step (Table 11), using different modalities (CT-MR) resulted in a slightly lower performance than using the same modality for registration (CT-CT and MR-MR).

*Table 11 – Image modality sensitivity analysis of lesion matching accuracy results, using automated dilation.*

| Registration Modalities | # of Matching Decisions | CL Acc. | NL/DL Acc | Overall Acc. |
|---|---|---|---|---|
| CT-CT | 199 | 0.89 | 1.00 | 0.94 |
| CT-MR | 640 | 0.80 | 0.99 | 0.90 |
| MR-MR | 114 | 0.91 | 1.00 | 0.94 |

#### *3.4.2.4 Registration Mode Sensitivity Analysis*

Table 12 shows the lesion matching accuracy results for the direct and indirect registration modes. The results for both registration modes were similar, with CL accuracy results that differed by 0.02, the same NL/DL accuracy, and overall accuracy that differed by only 0.01.

*Table 12 – Registration mode sensitivity analysis of lesion matching accuracy results, using automated dilation. Only considered patients with more than two images in their series (patients 1 and 2)*

| Registration Mode | # of Matching Decisions | CL Acc. | NL/DL Acc | Overall Acc. |
|---|---|---|---|---|
| Direct | 490 | 0.83 | 1.00 | 0.90 |
| Indirect | 463 | 0.81 | 1.00 | 0.89 |

#### *3.4.2.5 Track Size Sensitivity Analysis*

Figure 27 shows the identical track ratio as a function of numbers of layers (images) and number of nodes (lesion contours) in the tracks. Tracks with 4 layers were exclusively from patient 1 and tracks with 3 layers were exclusively from patient 2. Most lesion tracks had fewer than 5 lesions. There were 19 tracks with five or more nodes. This includes 10 tracks with 5 nodes, 3 tracks with 6 nodes, and 6 individual tracks each with 9, 10, 11, 12, 18, and 19 nodes.

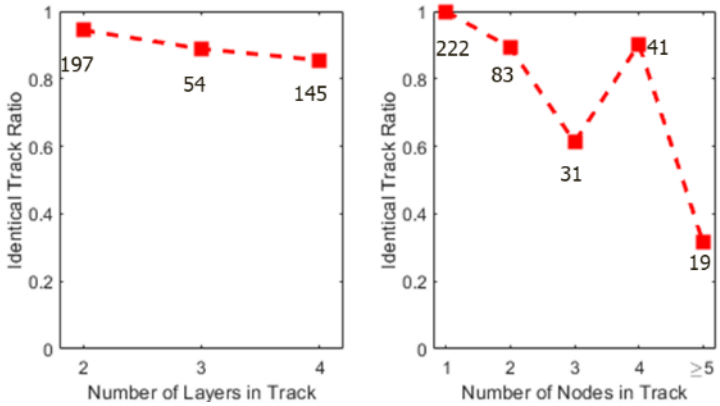


*Figure 27 – Track size sensitivity analysis of identical track ratio results. The numbers next to the markers indicate how many tracks were present in each category.*

#### *3.4.2.6 Failure Modes*

There were 108 errors out of 1,183 (9%) automated matching decisions made using automated dilation. The most common failure mode was in the registration (55/108 errors), where lesions were incorrectly superimposed due to inaccuracies in registration either by presenting a spurious overlap, or by failing to overlap when lesions should have. This failure mode was further divided into four sub modes (Table 13). The borderline sub mode was defined as lesions that wrongfully did not intersect but shared a border. The liver and spine registration failure sub modes denote that an inaccuracy in registration of these organs was the driver of the error. The second most common failure mode was inaccurate lesion clustering (32 errors), where incorrect lesion clustering resulted in a matching error. Less frequent failure modes were shoulder lesion (9 errors), which failed to intersect when the patient's arm positioning changed significantly between scans (e.g., patient imaged with arms up, then imaged with arms down). The resection failure mode (7 errors) happened exclusively for one patient (patient 2) that had the inferior portion of the liver resected between scans. In this case, the WDR wrongfully distorts the resected liver to match the unresected liver, resulting in matching errors. Finally, the least common failure mode was field of view (FOV) (5 errors), which happened in the edges of the imaging FOV where the registration had higher inaccuracies, impairing lesion intersection.

*Table 13 - Breakdown of the failure modes involved in the 108 errors made using automated dilation.*

| Failure Modes | # of Errors – n (%) |
|---|---|
| Registration | 55 (51%) |
| *Registration (borderline)* | *39 (36%)* |
| *Registration (liver)* | *6 (6%)* |
| *Registration (various)* | *4 (4%)* |
| *Registration (spine)* | *6 (6%)* |
| Cluster | 32 (30%) |
| Shoulder lesion | 9 (8%) |
| Resection | 7 (6%) |
| Field of view (FOV) | 5 (5%) |
| **Total** | **108** |

## 3.5 DISCUSSION

In this chapter we introduced a lesion tracking methodology that was validated on a population of neuroendocrine tumor patients with high-burden metastatic disease. The methodology is based on lesion contour superposition from registration of longitudinal multimodality images to the baseline image of an image series. Image registration was accurate, with a median residual landmark error of 2.6 mm for WDR (Figure 22) which is comparable to the landmark based residual errors presented in section 2 for whole-body PET/CT (Santoro-Fernandes et al., 2021) and for lung CT (Yin et al., 2011). We studied the lesion matching and tracking performances under various fixed and automated dilations, of which the latter resulted in the highest proportion of identical lesion tracks (90%) and overall accuracy (0.91) (Figure 24). The weak correlation ($\rho = 0.3$) between the overall accuracy (indicative of lesion matching performance) and the identical track ratio (indicative of lesion tracking performance) suggests a considerable degree of independence between the two metrics because multiple matching mistakes can be included in the same track or spread through various tracks. Lesion tracking accuracy variation by anatomical region was low. The lowest lesion matching accuracy was observed in the spine (0.88), and the lowest identical track ratio was observed in the arms (78%), with standard deviations of 0.03 for lesion matching accuracy and 8% for identical track ratio across anatomical regions (Table

10). This suggests that the methodology performance is robust regarding body-part. Use of inter-modality led to inferior registration accuracy (Figure 22) and lesion matching (Table 11), indicating a degree of correlation between registration and lesion matching performances, highlighting the importance of accurate registration as a prerequisite for accurate lesion matching.

The method introduced by Yan et al is a non-registration approach to lesion tracking that consists of evaluating the similarity between lesion features derived using deep learning methods (Yan et al., 2018). They performed whole-body lesion tracking limited to a single modality (CT). Their work is the only whole-body lesion matching work that reported on lesion matching accuracy. They reported precision and recall, which were respectively 0.86 and 0.92, leading to an F-score of 0.89. In comparison, our methodology had higher precision (0.91), recall (0.94), and F-score (0.92). These gains in precision and recall and minor and suggest the non-inferiority of the methodology introduced in this work. However, the methodology herein introduced has the additional advantage of handling multiple imaging modalities. In lieu of precision/recall, we chose the lesion matching accuracy (overall, CL, and DL/NL) as a performance metric for lesion matching because its interpretation is more intuitive. Furthermore, the precision-recall analysis used by Yan et al. considers lesion matches that aren't introduced by the automated tracking, since two lesion within the same track might not be a match (have a connecting edge).

In a recent work, Szeskin et al. introduced a simultaneous lesion tracking and segmentation approach that utilizes image registration followed by a convolutional neural network simultaneous segmentation approach. Their work is specific for liver lesions on CECT and was benchmarked on 50 scan-pairs containing 492 lesions. They report precision of 0.86 and recall of 0.90, leading to a F-score of 0.88. All lesion matching metrics were inferior when compared to the results of this chapter (0.91 precision, 0.94 recall, 0.92 F-score). In summary, our methodology achieved superior lesion matching results on a more strenous test, that of tracking lesions across multimodalities, multiple time-points, and spread through the whole-body.

The main challenge in this chapter was, just as in chapter 2, to determine the reference lesion matches and tracks, which is subject to unavoidable observer bias. Nevertheless, most of the lesion matching decision are conspicuous, and the decisions that elicit confusion would also be subject to inter-observer variability in a clinical setting. Therefore this challenge does not take away from the potential clinical utility of the developed methodology. Another challenge is the limited number of patients available for benchmarking the methodology. Despite only testing on 6 patients, many matching decisions (1,183) were available to validate the lesion matching performance. Furthermore, the standard deviation across patients for lesion matching accuracy was 0.07 and for identical track ratio was 9% (Table 11), which indicates a small outcome variation between patients. However, further investigations with more patients and different applications (imaging modalities, FOVs, cancer types, etc.) are necessary to evaluate the robustness of the methodology. A third challenge was the different arm positioning between PET/CT and PET/MR (above the head, to the side of the torso), which introduces anatomical discrepancies that cannot be reconciled by the deformable registration approach without unrealistic image distortion. To circumvent this limitation, the whole-body images were constrained to head, torso, and legs when needed (arm lesions were absent in these instances).

The use of the lesion tracking methodology in clinical applications should be cautious since an important drop in the identical track ratio was observed as the lesion tracks increased in number of layers and in number of nodes (Figure 27). This result is expected since identical track ratio is a very strict metric (a single error suffices to break the identicalness). This issue can be diminished in clinical applications by considering the most recent image series subset for patient evaluation, especially when the patient is surveilled over many years. More importantly, automated lesion tracking should serve as an aid to the physician, allowing for human intervention to correct errors in the lesion tracks. A work by Moltz et al. (Moltz et al., 2012) has shown that automated lesion matching can save time and reduce inter-reader variability of clinical decisions, suggesting that human correction can be faster than the infeasible task of manual lesion tracking. The feasibility of the human corrections is reliant on sufficient automated tracking accuracy, which this work has shown to be encouraging

(90% overall identical track ratio). Another factor that facilitates human correction of lesion tracks is that errors in lesion matching often lead to joined lesion tracks rather than completely incorrect decisions. This is mostly observed in anatomical regions of dense lesion population. Figure 28 illustrates a situation where three reference lesion tracks joined because of wrong automated lesion matches. In a clinical setting, providing that the implementation of the automated lesion tracking methodology allows for rapid and guided manual intervention, an operator could separate the three tracks in less time than individually matching all 15 lesion contours involved.

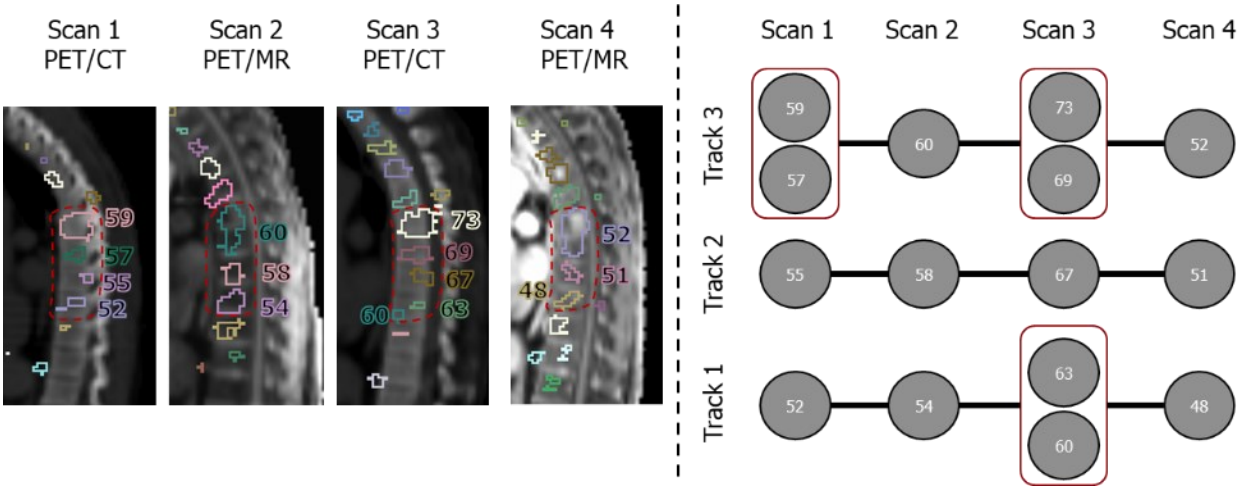


*Figure 28 – LHS: sagittal slices of the thoracic spine for each time-point. The PET signal was omitted to improve visibility of the anatomic structures. The red dashed outline indicates the clusters determined by the automated tracking methodology. RHS: section of the reference lesion graph containing the lesion tracks, reference clusters are indicated by a red contour around nodes. Due to matching errors, the three tracks shown in the figure were joined during automated tracking.*

Individual lesion tracking allows for the construction of lesion graphs (Figure 20), which are mathematical representations of the pathology identified in the image series of a metastatic cancer patient. The nodes of the graph represent the lesions and can encode information such as lesion volume, tracer uptake, grey levels, radiomics metrics, or any other quantitative information desired. The edges of the graph (lesion matches) establish the temporal relationship between these quantities, making for a complete tool in the quantitative imaging context. Graphs have been previously used as input for predictive models (Cangea et al., 2018; Duvenaud et al., 2015), using lesion graphs and lesion-level information as predictive metrics for patient

response to cancer treatment holds immense potential. Other future investigations include lesion tracking of images with intersecting but unequal FOVs, establishing the degree of certainty of the matchings to guide manual lesion track corrections, and automatically addressing different arms positionings.

Finally, our novel lesion tracking methodology not only performs with high accuracy, but also presents a few other advantages when compared to the 4D connectivity approach, which is the most commonly used and relies entirely on the lesion superposition caused by registration (Gerig et al., 2000; Kikinis et al., 1999; Köhler et al., 2019; Kuckertz et al., 2021; Metcalf et al., 1992). The first advantage is the automated lesion dilation (step 3), which increases the probability of establishing lesion correspondence without relying on registration accuracy alone. A second advantage is the ability to establish lesion tracks even when the lesion falls below the detectability limit in an intermediate time-point (intermittent lesion - Figure 17), which is possible in our methodology due to the combinatorial linear assignment (step 5). A third advantage is avoiding the wrongful joining of tracks caused by spurious overlap. Instead, the linear assignment (step 5) disregards spurious overlaps by globally optimizing lesion overlap volume, while lesion clustering (step 4) enables the matching of disjoint lesions. A fourth advantage is performing whole-body lesion tracking. Previous works have been limited to brain (Gerig et al., 2000; Kikinis et al., 1999; Köhler et al., 2019; Kuckertz et al., 2021; Metcalf et al., 1992; Shahar & Greenspan, 2005), liver (Kuckertz et al., 2022; Szeskin et al., 2023), or bone tissue (Yip & Jeraj, 2014). The fifth advantage is performing lesion tracking with multiple image modalities (PET/CT and PET/MR), this is the main addition to the methodology presented in section 2. Previous works that developed whole-body lesion matching were limited to PET/CT (Hering et al., 2021; Santoro-Fernandes et al., 2021) or CT images (Yan et al., 2018). In summary, the lesion tracking methodology developed in this chapter achieved highly accurate results and is the first registration-based whole-body methodology that does not present the 4D connectivity limitations and accepts multiple imaging modalities and multiple time-points as input. Other novelties of our methodology include accounting for lesion merging and splitting by considering lesion morphology and angular constraints, and an automated selection of the magnitude for the conformal lesion dilation to account for local variations in

lesion density. Finally, the introduced methodology is tracer independent since it assumes a pre-existing lesion contour map.

## 3.6 CONCLUSION

In this chapter we have developed a whole-body and multimodal lesion tracking methodology that enables lesion-level treatment response assessment of metastatic cancer patients. The developed methodology achieved high accuracy in lesion matching and tracking in a NET patient population imaged with $^{68}$Ga-DOTATATE PET/CT and PET/MR. The novel methodology introduced for tracking lesions through all time-points of an image series was made possible by using scan-pair-wise lesion matching decisions with a combinatorial approach that leads to lesion tracks. Our results are evidence that individual lesion tracking is feasible in a multimodality setting. However, the methodology benchmarking presented is limited and further validation is needed to identify the methodology's strengths and weaknesses and to consider clinical translation.

# 4 TREATMENT RESPONSE PREDICTION BASED ON COMPREHENSIVE, LESION-LEVEL, LONGITUDINAL DISEASE METRICS

In chapter 4, we apply the previously developed and validated methodology to build predictive models of peptide receptor radiotherapy (PRRT) outcomes in mNET patients. The predictive models are developed on the lesion-level and on the patient-level. Material from this chapter is adapted from a manuscript titled "Models using comprehensive, lesion-level, longitudinal PET-derived uptake features lead to superior prediction outcome in neuroendocrine tumor patients treated with $^{177}$Lu-DOTATATE" currently under preparation for submission to the European Journal of Nuclear Medicine and Molecular Imaging.

## 4.1 MOTIVATION

The treatment response of mNET patients treated with PRRT can vary significantly both on the lesion-level and on the patient-level. In the **lesion-level,** it is known as mixed response, heterogeneous response, or atypical response (Lyu et al., 2020). For these cases, introducing subtotal locoregional therapy targeting the persistent lesions could improve the overall clinical condition, leading to improved quality of life, meaningful disease burden reduction, delay of progression, and possibly to improved survival. However, currently, lesions are chosen at the discretion of the interventional radiologist based on technical factors, such as ease of access and proximity to vessels (Duffy et al., 2017). A methodology to **identify and select lesions** that are the optimal candidates to receive subtotal locoregional therapy is lacking.

In the **patient-level**, the response discrepancy in mNET patients treated with PRRT (section 1.4) renders predictive analysis of PRRT essential for personalized care of NET patients. If credibly informed with a patient-specific treatment outcome prediction, attending physicians could improve their decision-making regarding selection of patients for clinical trials, optimal monitoring frequency, optimal time for switching therapy, or for

combining different therapy strategies. One promising approach to stratify patients into good and poor responders is by using the standardized uptake value (SUV) of molecular somatostatin receptor imaging (SRI) (e.g., $^{68}$Ga-DOTATATE PET imaging). As described in detail in section 1.6, treatment Response prediction of mNET patients receiving PRRT has always been limited in one or more of the following ways: it did not take into account all of the lesions (non-comprehensive) (Ambrosini et al., 2015; Campana et al., 2010; Carlsen et al., 2020; Gabriel et al., 2009; Graf et al., 2020; Haug et al., 2010; Kratochwil et al., 2015; Öksüz et al., 2014; Werner et al., 2017, 2019; Zwirtz et al., 2022), it neglected longitudinal early response information (Carlsen et al., 2021; Ortega et al., 2021; Sharma et al., 2019; Urso et al., 2023), it was performed on the patient- rather than the lesion-level (Durmo et al., 2022; Ohlendorf et al., 2020; Opalińska et al., 2022; E. Pauwels et al., 2020). Moreover, SRI features are normally explored as single outcome predictors. A few works have suggested combination of features; Carlsen et al. suggested a combination of features through establishing cutoffs and grouping patients into classes (Carlsen et al., 2021), while Zwirtz et al. introduced a combination of features through multiplication of CT Hounsfield units and SRI features (Zwirtz et al., 2022). However, the combination of SRI features is in its infancy.

In summary, current efforts in **lesion-level** response prediction are incipient due to the technical difficulty in performing lesion-level assessment of response (chapters 2 and 3). On the **patient-level**, there is a possibility that comprehensive, lesion-level, longitudinal image analysis will improve treatment response prediction. In this chapter, we investigate the value of comprehensive, lesion-level, longitudinal imaging information to predict treatment response in the patient- and in the lesion-level. The treatment response prediction was investigated through mathematical models of long-term lesion persistence for the lesion-level, and progression-free survival for the patient-level.

## 4.2 Materials and Methods

### 4.2.1 Patient Population and Study Design

This multicentric retrospective study enrolled metastatic NET patients from the University of Wisconsin-Madison Carbone Cancer Center (UWCCC) and our collaborators at the Roswell Park Comprehensive Cancer Center (RPCCC). This study was reviewed and approved by the internal review board of both institutions and all ethics guidelines were followed. Patients that received their 1$^{st}$ cycle of treatment between March 2018 and December 2021 were included. All patients were treated with PRRT ($^{177}$Lu-DOTATATE), imaged with SRI ($^{68}$Ga-DOTATATE PET/CT or PET/MR) on the baseline (before 1$^{st}$ cycle of PRRT) and post-therapy (after last cycle of PRRT). A subset of the patients was also imaged in a third time-point, after post-therapy, named the long-term time-point.

For each patient a detailed chart review was performed, and the progression free survival (PFS) was determined as the time between the last PRRT cycle and the date of progression. Patients were followed until May 2023. Progression was determined by radiological progression (using RECIST 1.1), overall worsening of the clinical presentation, decision to change or perform additional treatment, or patient death.

### 4.2.2 Image Acquisition and Analyses

A diagram summarizing the image acquisition and analysis involved in this chapter is shown in Figure 29.

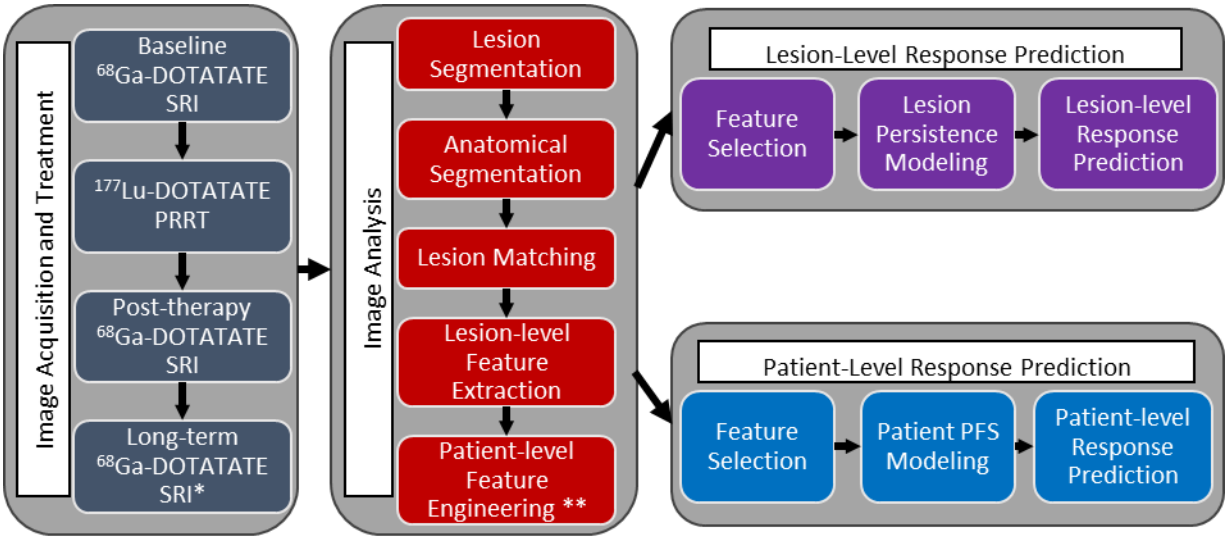


*Figure 29 - Diagram of chapter's 4 data acquisition and analysis. *The long-term images were used in the lesion-level response prediction only. **Patient-level feature engineering was used only for patient-level response prediction.*

#### *4.2.2.1 Image acquisition and Treatment*

Image acquisition followed standard clinical procedures. Patients fasted for 4 hours, and short acting somatostatin drugs were held for 24 hours before the scanning procedure. The administered activity was 0.054 mCi per kg of bodyweight of $^{68}$Ga-DOTATATE, with a lower limit of three and an upper limit of 5.4 mCi. The scanning started 60 minutes after tracer administration and the acquisition lasted 5 minutes per bed position. The standard treatment protocol was 4 cycles of $^{177}$Lu-DOTATATE PRRT (standard activity of 7.4 GBq (200 mCi) per cycle) administered eight weeks apart.

#### *4.2.2.2 Lesion segmentation*

The lesions present in the PET/MR images were segmented manually, similarly to the procedures described in sections 2 and 3. The lesions present in the PET/CT were automatically segmented using an in-house trained convolutional neural network (CNN). Despite the utilization of the lesion segmentation outputs, the CNN training was not performed as part of this dissertation. However, the training details are described here for a better understanding of the lesion segmentation used in subsequent response prediction steps. Two state-of-the-art 3D CNN architectures were employed for automated lesion delineation. The first was the *nnUNet*, which

follows a standard U-Net like architecture with automated hyperparameters inferred from input data feature and available compute infrastructure information (Isensee et al., 2021). Image patches of size 224×80×128 voxels were extracted for training and the loss function was set as the sum of Dice Coefficient and Cross Entropy. The model was trained for 1000 epochs using a batch size of 2 and 250 minibatches per epoch. Data augmentation was applied during training using the following techniques: rotation, scaling, gaussian noise, gaussian blur, brightness, contrast, low resolution simulation, gamma augmentation, and mirroring.

The second CNN architecture was the *Retina U-Net* (Jaeger et al., 2018). This model follows a feature-pyramid-like architecture for object detection with added high-resolution pyramid layers to encode semantic information, enhancing delineation accuracy. Image patches of size 128×128×128 voxels were extracted for training and the loss function was set as the sum of Dice Coefficient and Cross Entropy. The model was trained for 250 epochs with 600 minibatches per epoch and a batch size of 4. Random rotations and scaling were applied during training as data augmentation.

Each model was trained to delineate lesions (foreground) from background in a binary manner. The networks were trained on 116 images from 59 NET patients. 5-fold cross validation was utilized with identical train-test splits between the *nnUnet* and *Retina U-Net* models where 80% of the data was used for training and the remaining 20% was used for testing for each fold. Care was taken to ensure that scans of the same patient (i.e., follow-up imaging) were not distributed into training and testing groups for a given fold. In addition to assessing the performance of each method separately, we compiled ensembled delineation maps using the outputs of the *nnUNet* and *Retina U-Net* models via *Union* and *Intersection* ensembling. Thus, we assessed a total of four automated lesion delineation outputs: *nnUnet*, *Retina U-Net*, *Union Ensembling*, and *Intersection Ensembling*. The *nnUNet* and the *Retina U-Net* models were trained on a NVIDIA RTX Titan GPU and a NVIDIA GeForce RTX 3090 workstation, respectively, both with 25 GB of memory. Performance details are presented elsewhere (A. Weisman et al., 2022). The generated contours were manually reviewed and corrected under the

advice of a senior nuclear medicine physician. Lesions smaller than 1 $cm^3$ were disregarded to avoid partial-volume effects.

#### *4.2.2.3 Anatomical segmentation*

The body-part associated with each lesion was determined by registering segmentation atlases to all images. Eight body-parts were segmented using the atlas registration, namely head and neck, chest, lungs, abdomen, liver, spleen, thighs, arms, and spine. Using the atlas registration, the background liver and spleen uptakes were automatically extracted by selecting the statistical mode of the SUV distribution for each organ volume, these quantities were named $SUV_{liver}$ and $SUV_{splean}$. Normal background uptake normalization was shown to be useful in outcome prediction (Durmo et al., 2022; Haug et al., 2010; Opalińska et al., 2022; Ortega et al., 2021; E. Pauwels et al., 2020; Sharma et al., 2019; Urso et al., 2023). This approach is robust against registration inaccuracies, SUV spilling from other anatomical structures, and presence of lesions (tissue volumes more avid to somatostatin receptor within the organs). Details about the body-part labeling method were discussed in section 3.2.

#### *4.2.2.4 Lesion matching*

Following lesion segmentation, the novel method developed and validated in this work (chapters 2 and 3) was used to longitudinally track the lesions present across the baseline, the post-therapy, and the long-term images (when applicable) (Santoro-Fernandes et al., 2021). This algorithm accounts for lesion clustering that can be caused by lesion merging and splitting during treatment response. This process was extensively discussed in chapters 2 and 3.

#### *4.2.2.5 Lesion-level Feature Extraction*

Lesion-level quantitative features were extracted from each lesion contour. The following five features were extracted from each lesion (Table 14): $iSUV_{max}$, $iSUV_{min}$, $iSUV_{mean}$, $iSUV_{median,}$ $iSUV_{total}$, $iSUV_{vol}$, and $iSUV_{var}$, $iSUV_{skew}$, and $iSUV_{kurt}$ (The preceding "i" denotes the lesion-level nature of each SUV feature).

*Table 14 - Lesion-level features extracted. *Skewness and kurtosis were only used for lesion-level predictions*

| Feature Symbol | Meaning |
|---|---|
| $iSUV_{max}$ | SUV from the voxel with maximal uptake within lesion "$i$" |
| $iSUV_{min}$ | SUV from the voxel with minimal uptake within lesion "$i$" |
| $iSUV_{mean}$ | Mean uptake of all voxels within lesion "$i$" |
| $iSUV_{median}$ | Median uptake of all voxels within lesion "$i$" |
| $iSUV_{total}$ | Sum of the uptake of every voxel within lesion "$i$" |
| $iSUV_{vol}$ | Segmented volume of lesion "$i$" according to the functional imaging |
| $iSUV_{var}$ | Variance of the uptake of all voxels within lesion "$i$" |
| $iSUV_{skew}$* | Skewness of the uptake of all voxels within lesion "$i$" |
| $iSUV_{kurt}$* | Kurtosis of the uptake of all voxels within lesion "$i$" |

After lesion matching, longitudinal feature variation was calculated, namely $\Delta iSUV_{max}$, $\Delta iSUV_{min}$, $\Delta iSUV_{mean}$, $\Delta iSUV_{median}$, $\Delta iSUV_{total}$, $\Delta iSUV_{vol}$, $\Delta iSUV_{var}$, $\Delta iSUV_{skew}$, and $\Delta iSUV_{kurt}$ where

$$\Delta iSUV_j = iSUV_{j,post-T} - iSUV_{j,baseline} \tag{25}$$

and the index $j$ denotes each SUV feature. The relative feature variation was also measured as

$$\Delta iSUV_{j_{rel}} = \frac{iSUV_{j,post-T} - iSUV_{j,baseline}}{iSUV_{j,baseline}} \tag{26}$$

The inter-correlation between each lesion-level feature relative variations in the three-image time-point subset was calculated to assess the degree of independence between the features.

#### *4.2.2.6 Lesion-Level feature Change Assessment*

Lesion matching allowed for determining the longitudinal feature variations. Using the matching information, lesions were classified into favorably responding or unfavorably responding response categories. These response categories were based on the feature change from baseline to post-therapy images. Favorably responding lesions were the ones with feature change $\leq$ 30% and disappearing lesions. Unfavorably responding lesions were those with feature change >30% and new lesions. The rationale for this classification was that the lesions showing an increase in uptake despite PRRT administration are likely to be persistent lesions. The threshold in increase of 30% was selected based on similar findings for other PET tracers (Lin et al., 2016). Based

on their identification in the long-term image, the lesions were further classified as disappearing lesions (DL) or persistent lesions (PL). A PL lesion was identified and segmented on the long-term PET image. A DL lesion was not present in the long-term image. This later lesion classification was only performed in the subset of patients that were imaged on three time-points since a long-term image was necessary.

#### *4.2.2.7 Patient-Level Feature Engineering*

Patient-level features were engineered by calculating seven statistics across all lesions: count, sum, average, median, minimum, maximum, and variance. Furthermore, the features were used with and without liver or spleen background mean uptake *normalization*. The patient-level feature engineering process was repeated for 4 variations of lesion-level features: (1) at the baseline scan, (2) at the post-therapy scan, (3) their *longitudinal* variation, and (4) their relative longitudinal variation. Additionally, this process was performed for subsets of lesions separated by their *anatomic location* (AL) and by their *persistence status* (DL, NL, PL). To illustrate this process, Figure 30 shows the engineering of one example patient-level feature.

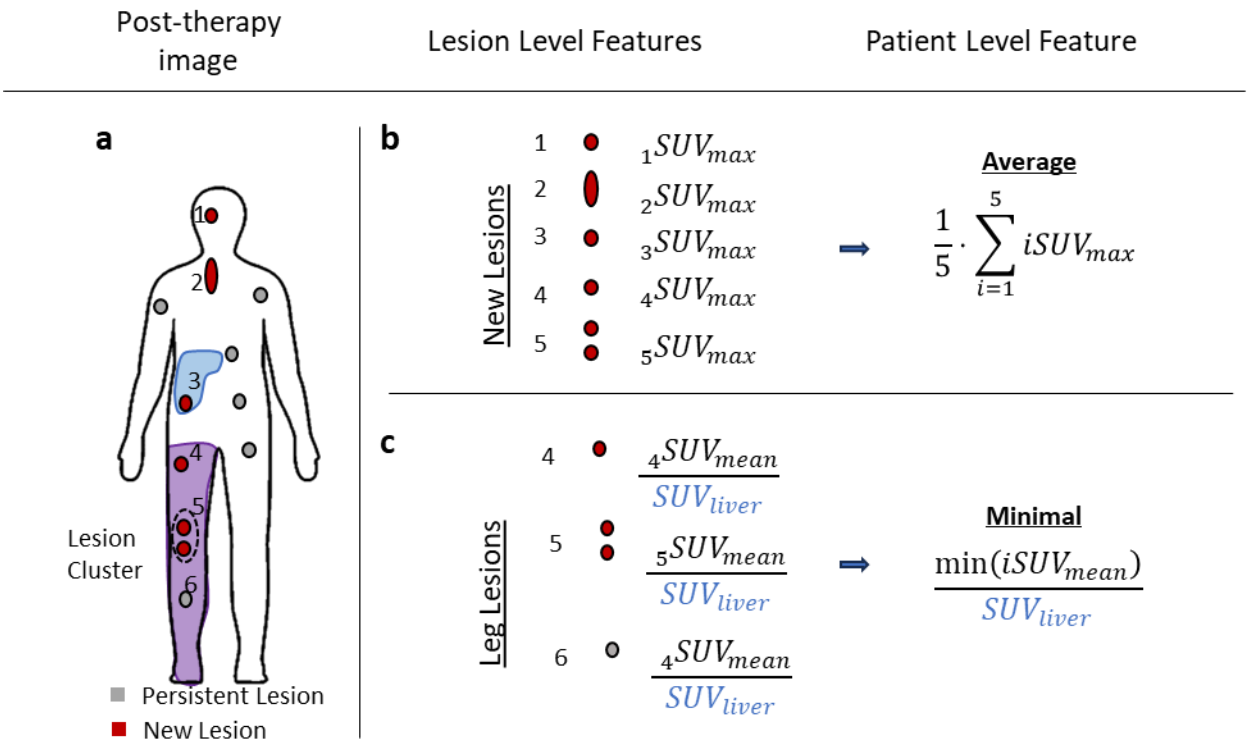


*Figure 30 – (a) Example of the patient-level feature engineering. In this example the engineered features were (b) the average $iSUV_{max}$ of the new lesions at the post-therapy scan, and (c) the minimal $iSUV_{mean}$ of lesions located in the leg normalized to the background liver uptake.*

## 4.2.3 Lesion-Level Response Prediction

### *4.2.3.1 Lesion Response Modeling*

A random forest (RF) model was trained to classify each lesion into (PL) or (DL) categories. The input to the RF model were longitudinal feature pairs. A longitudinal feature pair consists of a pair of lesion-level feature values (Table 14) for the same lesion, one value extracted from the baseline image, and one extracted from the post-therapy image. Only the lesions present (i.e., identified, and segmented) at the baseline or at the post-therapy images were used for the lesion-level response prediction analysis. The RF model was implemented in MATLAB (The MathWorks Inc.) version 2019a and was trained using data from the baseline and the post-therapy images. The lesions identification in the long-term image was used as the reference classification criterion of lesion-level response. The RF model output was a per-lesion binary classification of persistent lesion (PL) or disappearing lesion (DL). To prevent overfitting, the final model consisted of an ensemble of 30 decision trees,

and 5-fold cross validation was used in the model training process. In training the RF model, two sensitivity analyses were performed for the tuning of hyperparameters and of selected features: (1) we varied the number of allowed splits per tree from one to ten, (2) we varied the number of longitudinal feature pairs used as model input from one to nine (all available feature pairs). The feature selection process was performed iteratively by selecting in each iteration the feature whose addition would maximize the model's area under the receiver operating curve (AUC).

RF model performance was evaluated by measuring its AUC. The AUC confidence intervals were estimated using the bootstrap method. The models' AUCs were compared and the optimal model was defined by AUC maximization. For the optimal model, a detailed performance analysis was done by calculating its accuracy, specificity, and sensitivity at the optimal operating point determined via the Youden criterion.

#### *4.2.3.2 Longitudinal Sensitivity Analysis*

To assess the impact of longitudinal lesion-matching in the lesion persistence prediction performance, RF models were trained by neglecting either the baseline PET features or the post-therapy PET features. These models were compared to the optimal RF model.

#### *4.2.3.3 Lesion Classification Accuracy*

Lesion classification into PL or DL was performed in two ways: using the RF model, and using the lesion-level response categories (favorably or unfavorably responding). The accuracy of these two ways were compared using the Wilcoxon signed-rank test with significance value of 0.05 before correction by the Bonferroni method (0.006 after correction for 9 tests).

### 4.2.4 Patient-Level Response Prediction

#### *4.2.4.1 Feature Selection and PFS Modeling*

The feature engineering step of section 4.2.2.7 generates features which can be highly correlated. A technique for maximal relevance, minimal redundancy, and maximal robustness was used to select features,

similar to that used by Peng et al. (Peng et al., 2005). To select for relevance, the correlation of all the features to PFS was calculated (maximal relevance). For robustness, the correlations were evaluated using 5-fold cross validation and the average correlation was used for the analysis. The feature with the highest correlation was selected, and all features that had an absolute correlation coefficient to the selected feature above a threshold of 0.5 were removed from the pool of possible features (minimal redundancy). After which the remaining feature with the highest correlation to the outcome was selected. This process was iteratively repeated until the ten most relevant features were determined. The 0.5 threshold is a parameter of the feature selection method and can be tuned to balance model accuracy (lower threshold) vs. robustness (higher threshold). The most relevant patient-level features were then used to build classical multivariate linear regression (MLR) models of predicted PFS (denoted as $\widehat{PFS}$). The models were fitted using a leave-one-out cross validation approach to prevent overfitting. Each model fitted in each fold of the leave-one-out cross validation process used the same set of selected features.

The number of features used in the models was varied from one to ten to determine an optimal number of features ($N_{opt}$) that maximized the model's concordance index (C-index). This quantity is commonly used in survival analysis and measures the correlation between the observed PFS and $\widehat{PFS}$ taking the censored observations into account. The $N$ optimal features were used in each iteration. To avoid overfitting, the $N_{opt}$ was selected considering a compromise between performance and number of patients per feature (Vittinghoff & McCulloch, 2007).

#### *4.2.4.2 Model Performance Assessment*

The modeling performance was assessed by evaluating the C-index. The C-index uncertainty was derived from the binomial distribution standard deviation. A classification study was performed to assess the ability of the predicted $\widehat{PFS}$ to distinguish between good and poor responders. The patient population was stratified into good and poor responders and a receiver operating characteristics (ROC) analysis was used to evaluate the

classification performance and an optimal cutoff for $\widehat{PFS}$ was determined by maximizing the Youden J statistic. In the ROC analysis, the observed PFS served as the ground truth for classification into poor and good responders according to a threshold determined empirically based on the observed PFS distribution, while $\widehat{PFS}$ was the predictive variable whose optimal cutoff was determined. Bootstrapping was used to evaluate the uncertainty of the ROC analysis outcomes. The $\widehat{PFS}$ was used as predictor for a Kaplan-Meyer survival analysis, and the log-rank test was used to evaluate the significance of the stratification. In the Kaplan-Meier analysis, the optimal cutoff was the one that minimized the log-rank p-value between stratified groups. Additionally, univariate Cox Proportional Hazards Model analysis was performed.

#### *4.2.4.3 Model Sensitivity to Selected Features*

Several sensitivity analyses were performed to probe the robustness of the outcomes to different aspects of the employed methodology. In all of them, the optimal number of features ($N_{opt}$) identified was used. The robustness of feature selection was evaluated by using *alternative features* to fit the MLR model. The four most correlated *alternative features* were associated with each $N_{opt}$ selected feature. The most correlated alternative features were the ones that had the highest correlation to the original selected feature. The MLR models were fitted using alternative feature sets and the C-Index of the prediction was evaluated as a function of the *feature variation degree*. This is a measure of the distance between the used set of alternative features and the originally selected ones. For example, a feature variation degree of 0 means that the model was trained with the original features. A feature variation degree of 1 means that one of the original features was varied to its correspondent first most correlated alternative feature. A feature variation degree of 2 means that two original features were varied to their correspondent first most correlated alternative feature or that one original feature was varied to the second most correlated alternative feature, etc. Every possible combination for each feature variation degree was assessed and the median C-index was calculated.

#### *4.2.4.4 Model Sensitivity to Image Analysis*

The sensitivity of the prediction results to four aspects of image acquisition and analysis was investigated. First, the *longitudinal information* sensitivity analysis simulated the lack of post-therapy images. Therefore, only features from the baseline were considered for modeling. Second, the *comprehensiveness* sensitivity analysis simulated using only the five most prominent lesion contours (highest $SUV_{total}$). We have chosen to evaluate five lesions since it is a common number for non-comprehensive studies, justified by RECIST1.1 and PERCIST guidelines. Third, the *anatomical segmentation* sensitivity analysis simulated the lack of body-part labels and liver and spleen background uptake normalizations. Fourth, the *lesion matching* sensitivity analysis simulated the lack of lesion matching, therefore the lesion-level uptake variation and the persistence status of the lesions were not considered, however, patient-level uptake variation was still available.

#### *4.2.4.5 Benchmarking*

Significant predictor image features were identified from the literature and were used to benchmark the prediction performance. Features previously identified as predictors in the literature were disregarded if the reported significance would not pass a Bonferroni corrected hypothesis test. These features were directly used in Kaplan-Meier and Cox proportional hazards ratios survival analyses. The results were compared to MLR PFS modeling using the C-index, AUC, and hazard ratios (HR). For the latter, the values were normalized to the maximal range value so that HRs are intercomparable.

## 4.3 RESULTS

### 4.3.1 Patient Population and Study Design

In total, 36 patients were enrolled in this retrospective study, details about the patient population are shown in Table 15. All patients underwent 1 to 4 cycles of PRRT (standard activity of 11.1 GBq (300 mCi) per cycle). Reasons for fewer than 4 cycles (10 of 36 patients) were the occurrence of adverse events, death, or

patient election to discontinue treatment. The median time between baseline image acquisition and start of therapy was 10 weeks (IQR = 12). The median time between the last cycle of therapy and the post-therapy image acquisition was 13 weeks (IQR = 13). The median observed PFS was 19 months (IQR=30). Details about imaging and treatment timings are shown in Table 16. The patient population was stratified into poor (PFS < 25 mo) and good (PFS ≥ 25 mo) responders. The stratification (Figure 31) was determined empirically based on a prominent ten-month discontinuity on the progression free survival trend. For ten of the 36 patients the progression observation was censored (28%). Only one of the ten censored patients was classified in the poor responder group.

*Table 15 – Patient demographics, disease, and treatment characteristics.*

| Parameter | Value | N Patients | Percentage |
|---|---|---|---|
| Number of patients | | 36 | |
| Institution | UWCCC | 31 | 86% |
| | RPCCC | 5 | 14% |
| Sex | Male | 15 | 42% |
| | Female | 21 | 58% |
| Age | <50 | 4 | 11% |
| | 50 to 60 | 10 | 28% |
| | ≥60 | 22 | 61% |
| Number of PRRT cycles | 1 | 4 | 11% |
| | 2 | 2 | 6% |
| | 3 | 4 | 11% |
| | 4 | 26 | 72% |
| Primary Site | Adrenal | 2 | 6% |
| | Liver | 2 | 6% |
| | Lung | 1 | 3% |
| | Pancreas | 6 | 17% |
| | Colon | 2 | 6% |
| | Small Bowell | 21 | 58% |
| | Thymus | 1 | 3% |
| | Unknown | 1 | 3% |
| WHO Grade | G1 | 17 | 47% |
| | G2 | 12 | 33% |
| | G3 | 4 | 11% |
| | Unknown | 3 | 8% |
| Previous Treatment | Somatostatin Analog | 36 | 100% |
| | Resection | 21 | 58% |
| | Chemotherapy | 13 | 36% |
| | Radiation | 10 | 28% |
| | Embolization | 9 | 25% |
| | Everolimus | 4 | 11% |
| | Other | 8 | 22% |

*Table 16 – Timing details of study design*

| Time interval | Median | 1st Quartile | 3rd Quartile |
|---|---|---|---|
| Baseline PET to 1st Cycle (wk) | 10 | 5 | 17 |
| Last cycle to Response PET (wk) | 13 | 8 | 21 |
| Total follow-up observation time (mo) | 48 | 33 | 59 |
| Progression free survival (mo) | 19 | 9 | 39 |

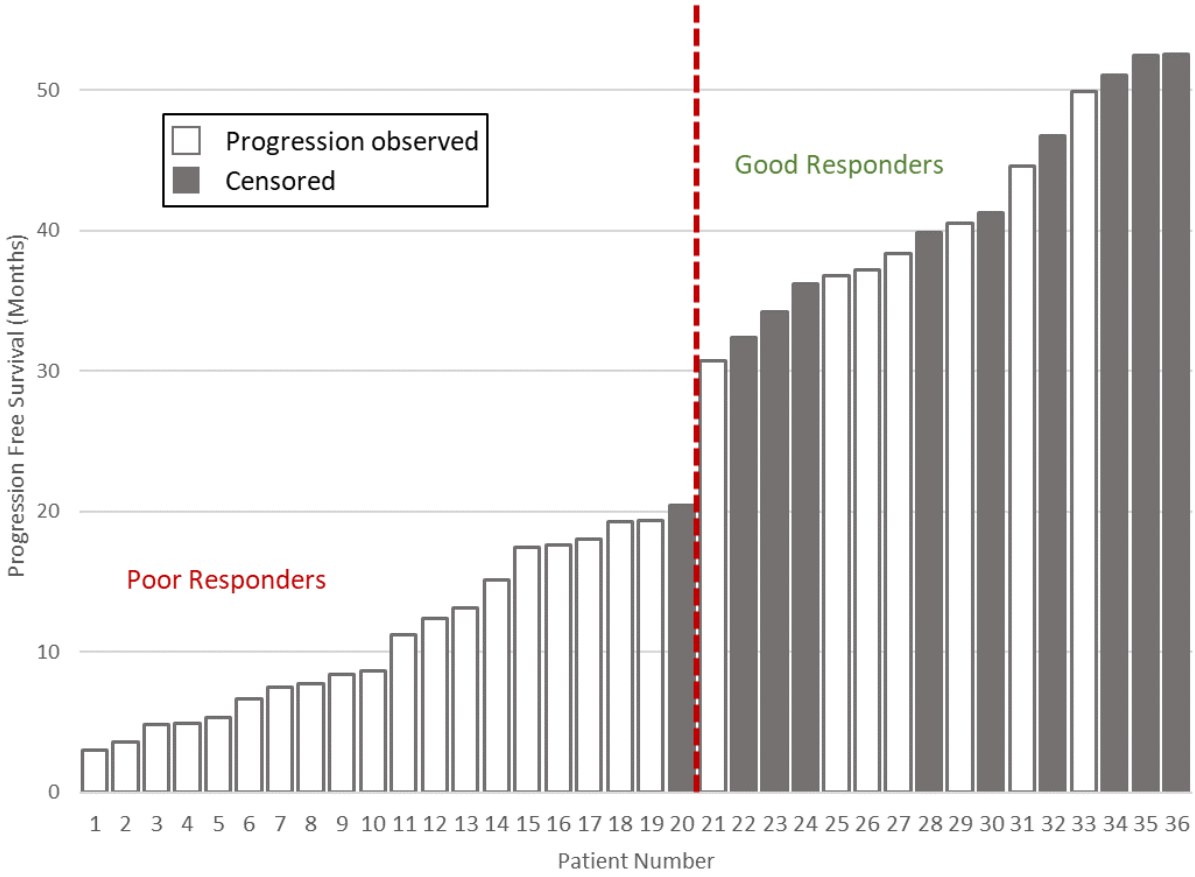


*Figure 31 – Progression Free Survival evaluated for each patient. Vertical red dashed line marks the cutoff point dividing poor responders (to the left), and good responders (to the right). Censored data is denoted by solid filled bars. Patients are numbered from worst to best response to PRRT.*

#### *4.3.1.1 Lesion-Level Response Prediction Subset*

Thirteen of the 36 patients were imaged at three time-points. This patient subset demographic information, primary site, and WHO grade are shown in Table 17 and the timing between the images and treatment delivery is shown in Table 18 and Figure 32.

*Table 17 – Details of the patient population subset used for lesion-level response prediction.*

| Parameter | **Value** | **# of patients** | **Percentage** |
|---|---|---|---|
| Number of patients | | 13 | |
| Sex | Male | 6 | 46% |
| | Female | 7 | 54% |
| Age | <50 | 2 | 15% |
| | 50 to 60 | 3 | 23% |
| | ≥60 | 8 | 62% |
| Number of PRRT cycles | 2 | 1 | 8% |
| | 3 | 2 | 15% |
| | 4 | 10 | 77% |
| Primary Site | Ileum | 5 | 38% |
| | Pancreas | 2 | 15% |
| | Adrenal | 1 | 8% |
| | Right colon | 1 | 8% |
| | Small Bowell | 4 | 31% |
| WHO Grade | G1 | 8 | 62% |
| | G2 | 4 | 31% |
| | Unknown | 1 | 8% |

*Table 18 – Timing details between imaging and treatment.*

| **Time interval** | **Median** | **1st quartile** | **3rd quartile** |
|---|---|---|---|
| Baseline PET to 1st Cycle (wk) | 10 | 6 | 14 |
| Last Cycle to post-therapy PET (wk) | 16 | 10 | 20 |
| Last Cycle to long-term PET (wk) | 65 | 52 | 72 |

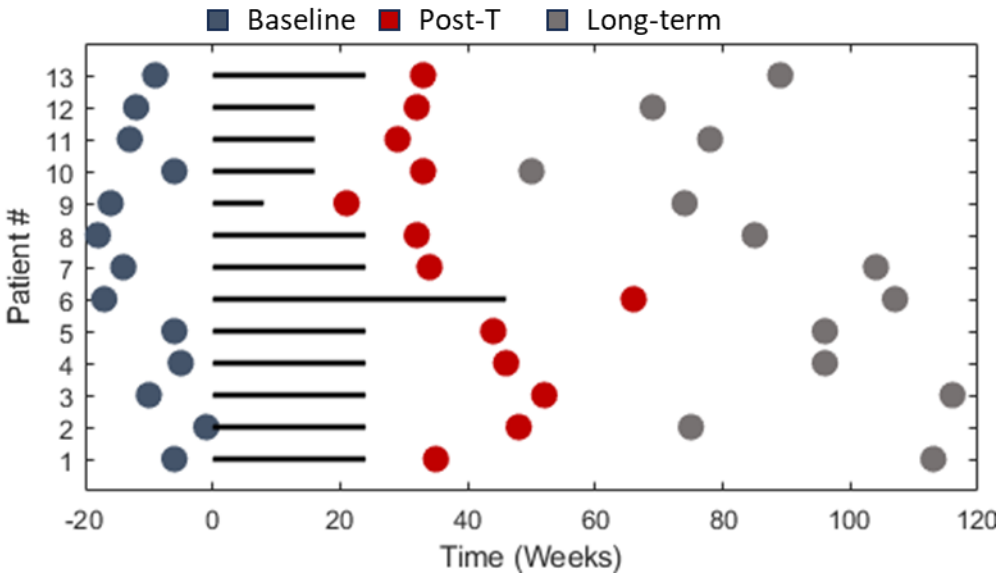


*Figure 32 – Time of acquisition of the somatostatin receptor images. The black continuous line represents the duration of PRRT treatment.*

### 4.3.2 Image Acquisition and Analyses

Eighty-five PET/CT and PET/MR scans were obtained in this study, two for each of the 36 patients and one additional image for each of the 13 patients with long-term image. Seventy-nine scans were PET/CT (92%), with only six (7%) PET/MR scans (Table 19).

*Table 19 - Image acquisition scanner information.*

| Parameter | Value | Number | Percentage |
|---|---|---|---|
| Number of Scans | Total | 85 | |
| Per Scanner | GE - Discovery 710 | 19 | 22% |
| | GE - Discovery IQ | 10 | 12% |
| | GE - Discovery MI | 43 | 51% |
| | GE - Signa (PET/MR) | 6 | 7% |
| | GE – Discovery ST | 1 | 1% |
| | Philips - Vereos | 4 | 5% |
| | Siemens - Biograph mCT | 2 | 2% |

Considering the baseline and post-therapy, 2307 lesion tracks were identified in the patient population. The median number of lesions identified per patient was 35, with a minimum of 3 lesions and a maximum of 328

lesions (IQR = 66). More image acquisition and lesion-specific information is shown in Table 20. An example of maximum intensity projections of the PET/CT images is shown in Figure 33, where the lesion and anatomical segmentations can be observed.

*Table 20 –Lesion-level image analysis characteristics.*

| Parameter | Value | Number | Percentage |
|---|---|---|---|
| Number of lesion tracks | Total | 2307 | |
| Per Body-part | Chest | 386 | 17% |
| | Abdomen | 441 | 19% |
| | Arms | 117 | 5% |
| | Head | 104 | 5% |
| | Pelvis | 254 | 11% |
| | Legs | 95 | 4% |
| | Liver | 433 | 19% |
| | Lung | 176 | 8% |
| | Spine | 301 | 13% |
| Per persistence status | Persistent lesion | 880 | 38% |
| | New lesion | 714 | 31% |
| | Disappearing lesion | 713 | 31% |

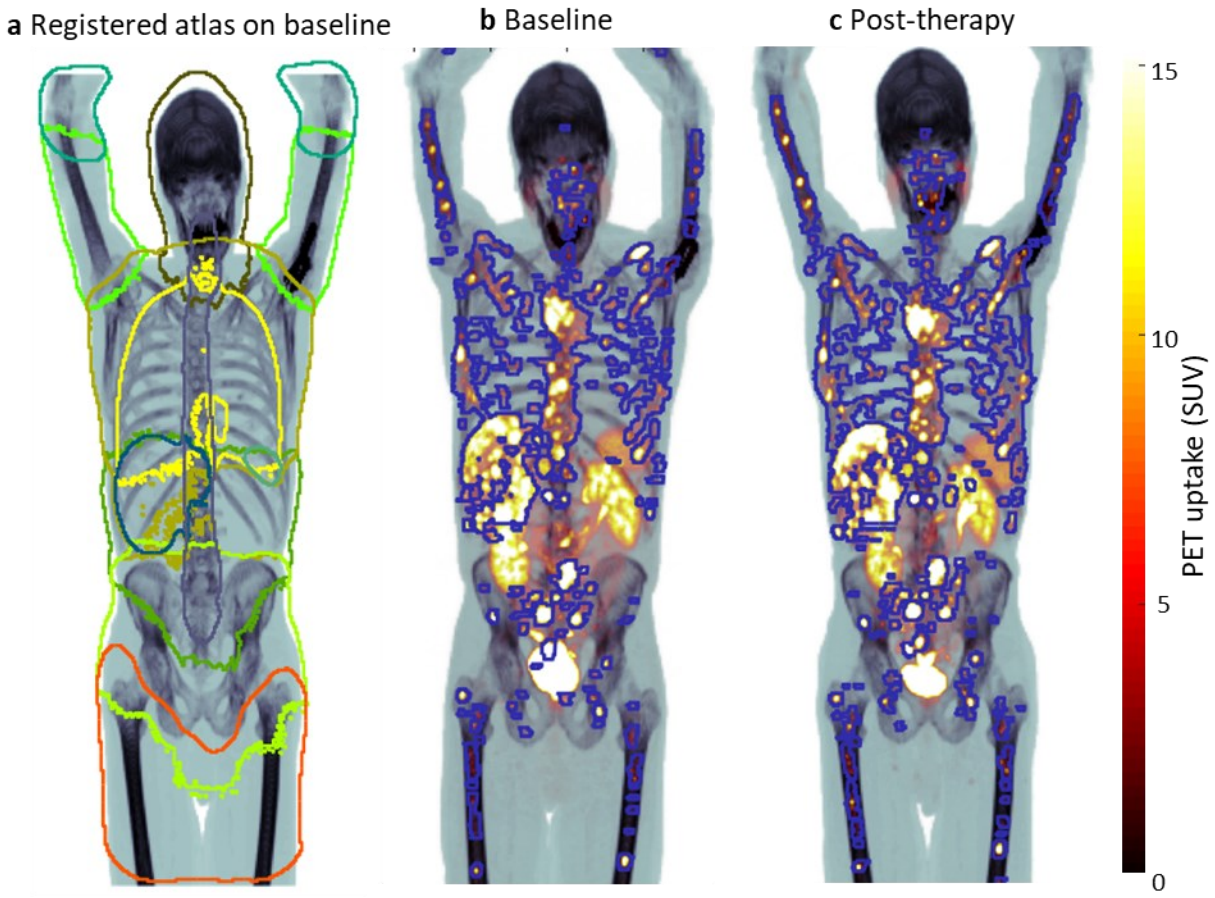


*Figure 33 – Example patient: a 73-year-old male presenting with 162 unique lesion tracks. This patient progressed 18 months after the 1st cycle of PRRT. The images are a) a maximum intensity projection (MIP) of the CT and the body-part segmentation atlas, b) along with a MIP of the PET/CT fusion with the lesion mask superimposed for the baseline PET/CT acquisition and c) for the post-therapy acquisition.*

In the subset of patients with three imaging time-points, 39 images were analyzed, three per patient (Table 21). The image analysis process identified and contoured 632 lesions. The median number of lesions per patient was 33, with a minimum of 3 and maximum of 198 lesions per patient. Two of the 39 scans were PET/MRs. Of the 632 lesions, 110 were identified on the long-term image only. Therefore, 522 lesions were available for lesion response modeling.

*Table 21 – Imaging information and lesion-level patient information.*

| Parameter | Value | Number | Percentage |
|---|---|---|---|
| Number of Scans | Total | 39 | |
| Number of lesions | Total | 632 | |
| Per Body-part | Chest | 109 | 17% |
| | Abdomen | 84 | 13% |
| | Arms | 32 | 5% |
| | Head | 28 | 4% |
| | Pelvis | 69 | 11% |
| | Legs | 15 | 2% |
| | Liver | 121 | 19% |
| | Lung | 38 | 6% |
| | Spine | 136 | 22% |

The correlation between the relative variation of each lesion-level feature between baseline and post-therapy images is shown in Figure 34.

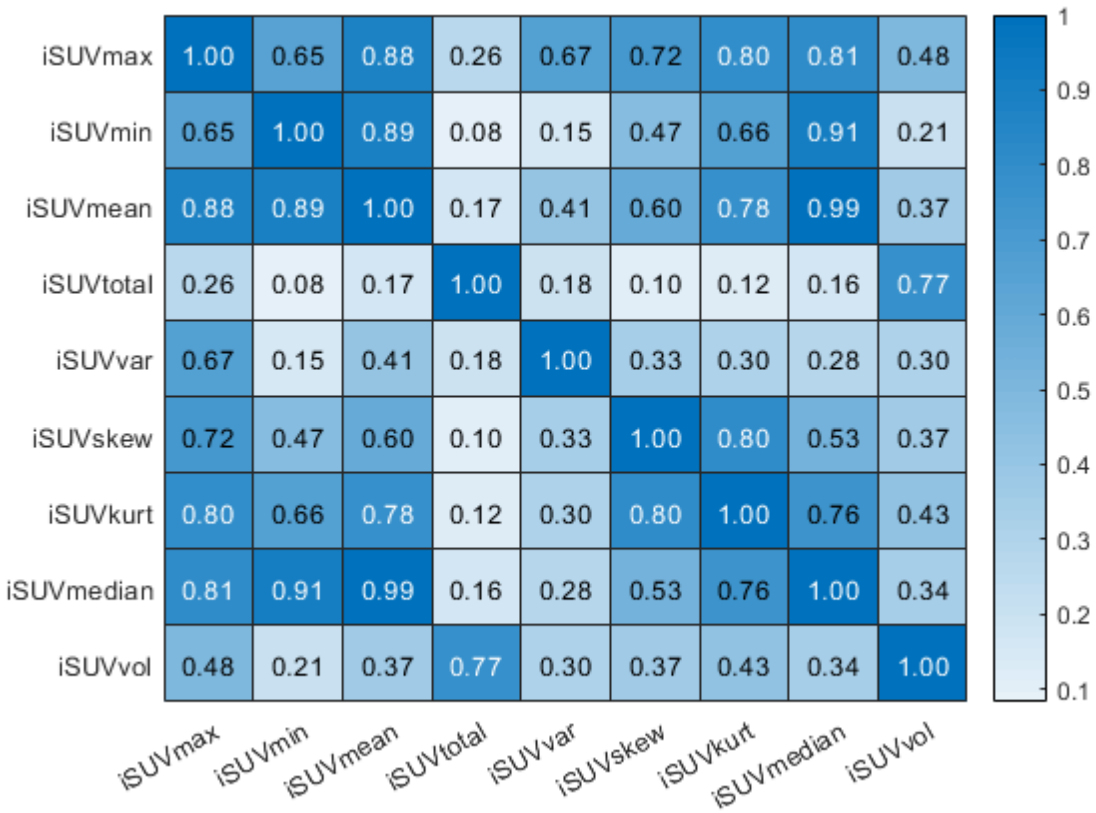


*Figure 34 – Correlation between the extracted lesion-level PET features*

### 4.3.3 Lesion-Level Response Prediction

#### *4.3.3.1 Lesion Response Modeling*

The number of longitudinal feature pairs used in the RF model had no significant impact in the model's AUC (Figure 35). The AUC fluctuates randomly and mildly. Therefore, we decided to use the simplest possible approach, including one single longitudinal feature pair to train the RF model.

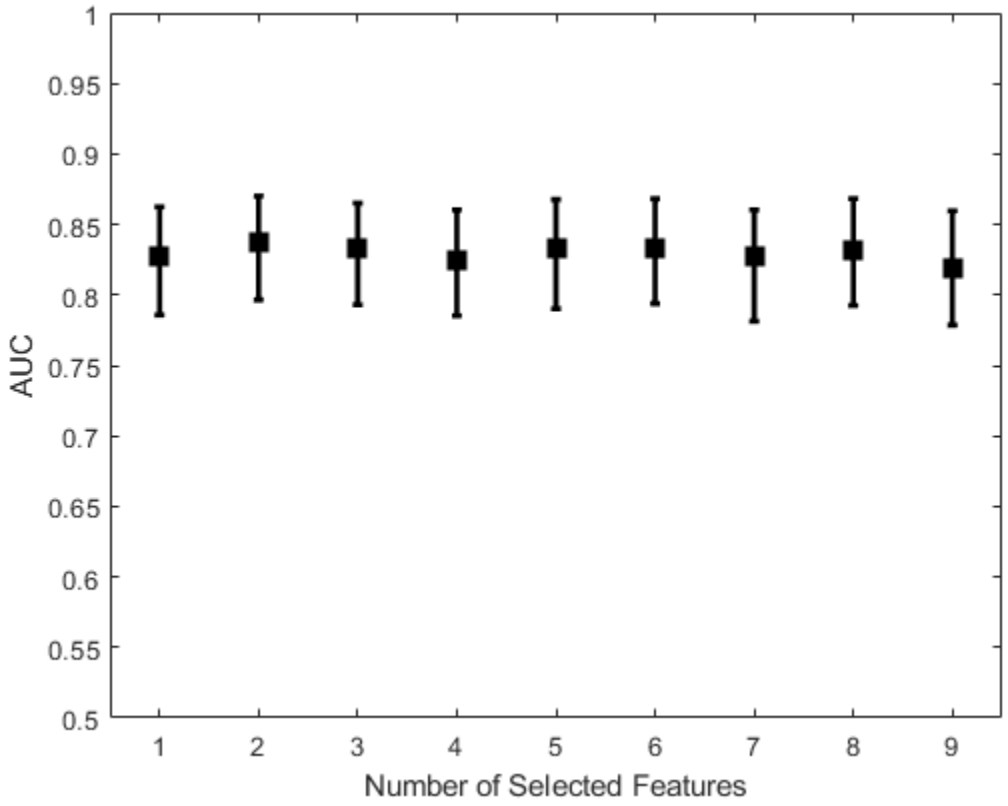


*Figure 35 – Random Forest model performance as a function of the number of selected longitudinal feature pairs for training. The error bars represent the 95% confidence interval.*

To decide which longitudinal feature pair to use, we studied the model's behavior when trained with each one (Figure 36). The AUC variation across lesion feature used to train the RF model was mild. The minimal AUC value was 0.76 for $iSUV_{skew}$ and the maximal AUC was 0.83 for $iSUV_{median}$. Considering the statistical equivalence between the longitudinal feature pairs, we have selected $iSUV_{max}$ to be used due to its straightforward interpretation and common use in the literature.

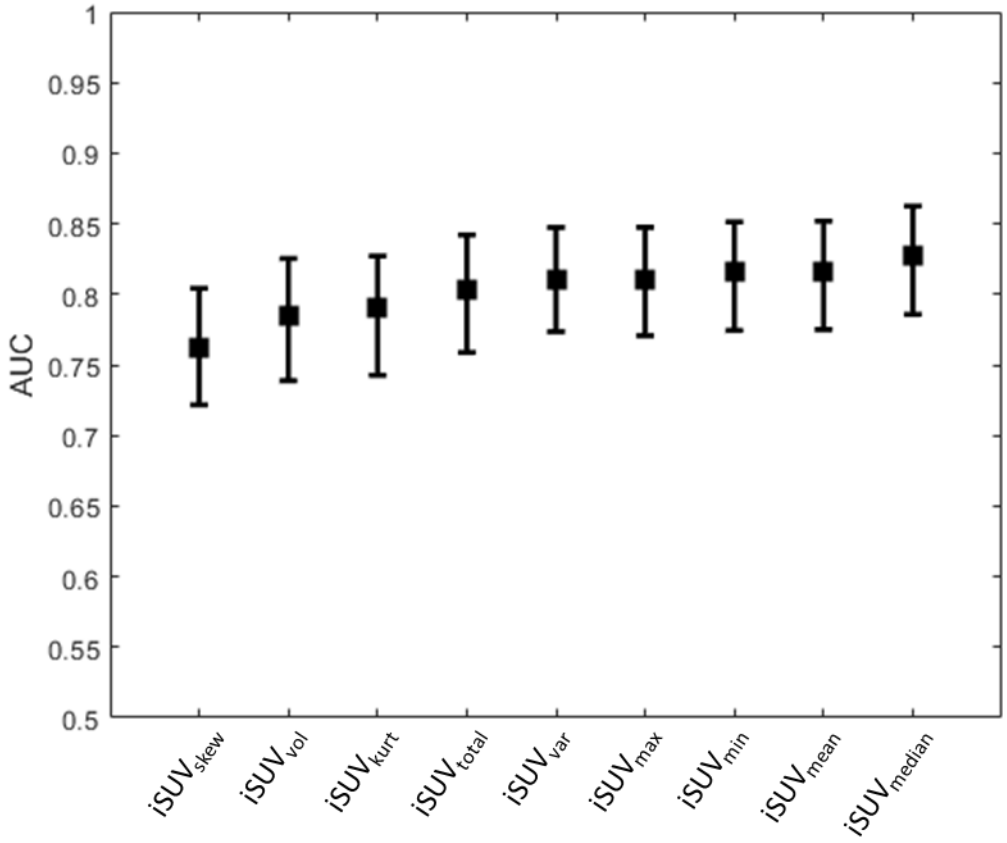


*Figure 36 – Performance (AUC) for a Random Forest model trained with a single longitudinal feature pair as a function of the chosen feature pair. The error bars represent the 95% confidence interval.*

Using the $iSUV_{max}$ as the input longitudinal feature pair, we have determined the optimal number of splits in the RF model by maximizing its AUC. AUC trends up slightly as the number of splits increases. The model AUC increased from 0.72 for one split to 0.82 at four splits, after which it plateaued (Figure 37). Therefore, the RF models used in this work were trained with 4 splits.

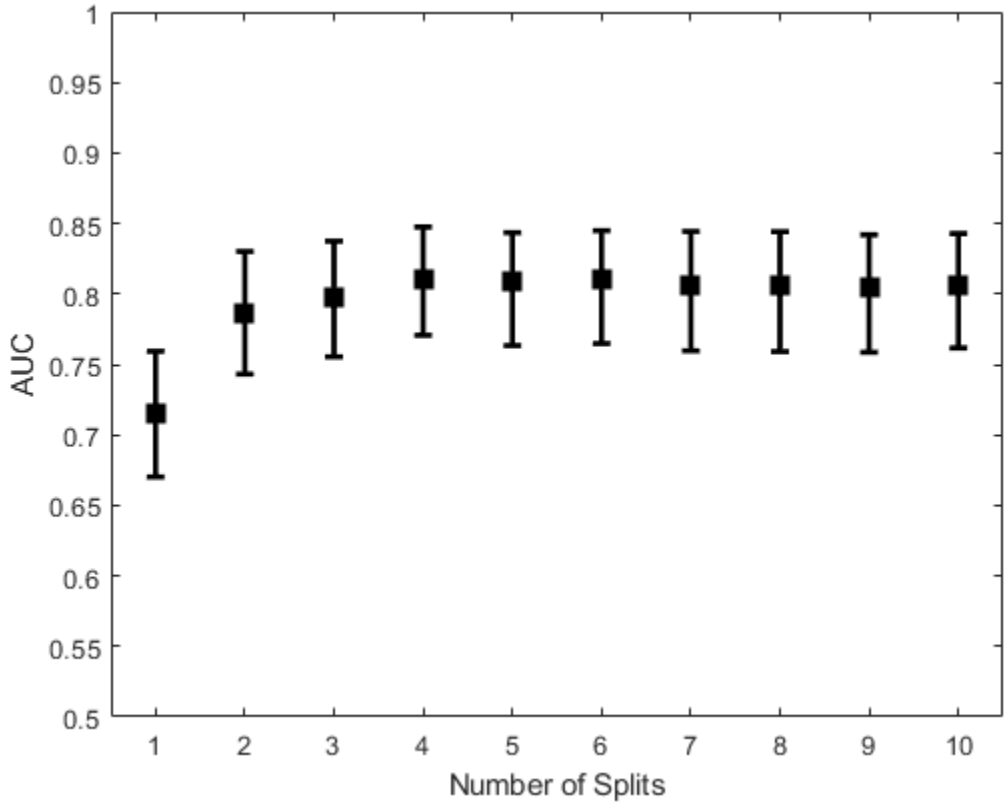


*Figure 37 – Random Forest model performance (AUC) as a function of the maximal number of splits allowed. The longitudinal feature pair used was the $iSUV_{max}$. The error bars represent the 95% confidence interval.*

The detailed performance analysis for the RF model trained using the $iSUV_{max}$ as the longitudinal feature pair with 4 maximal possible splits is shown in Figure 38. We have observed an AUC of 0.82, accuracy of 0.78 with specificity of 0.80 and sensitivity of 0.75. The box-plot for the RF classification scores is also displayed in Figure 38 showing the separation between the PL and the DL groups.

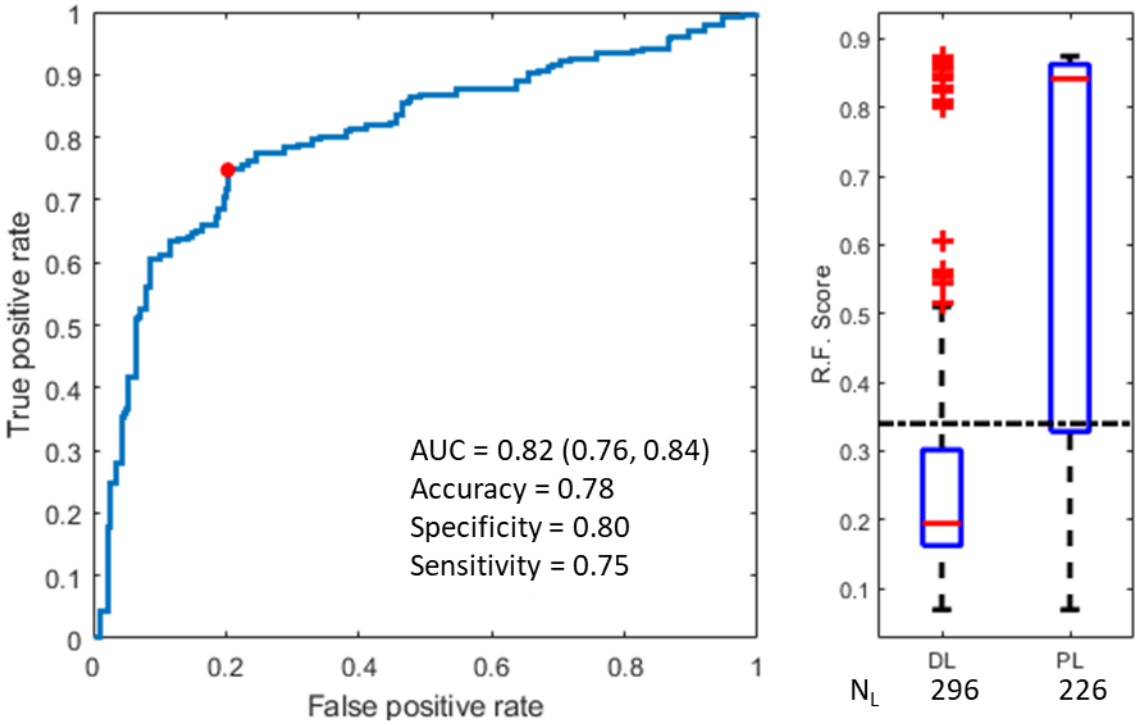


*Figure 38 – Detailed performance of the random forest model trained using the $iSUV_{max}$ longitudinal feature pair with 4 maximal splits. Left-hand side shows the ROC, and the right-hand side shows the boxplot of the model stratification between the two classes, namely persistent lesions (PL) and disappearing lesions (DL). The horizontal dashed line (right) and red dot (left) represent the optimal ROC cutoff.*

Figure 39 illustrates one of the 30 trees in the trained ensemble of the model described in in Figure 38.

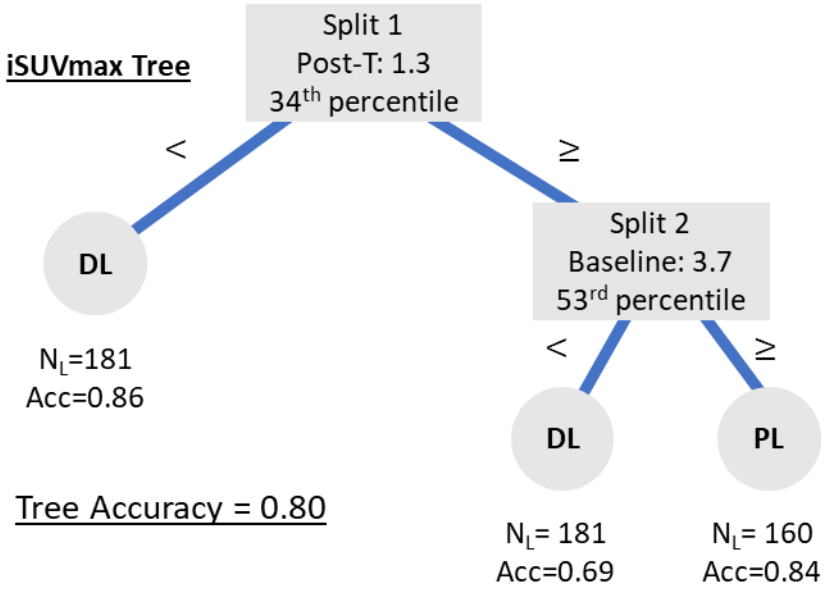


*Figure 39 – Example of one decision tree of the Random Forest model using $iSUV_{max}$ as the input variable. The final classes are persistent lesion (PL) and disappearing lesion (DL).*

#### *4.3.3.2 Longitudinal Sensitivity Analysis*

The longitudinal information sensitivity analysis results show that the prediction performance of the RF model was highly dependent on the longitudinal information (Figure 40). When the $iSUV_{max}$ longitudinal feature pair was used to train the model, an AUC of 0.86 was achieved, significantly higher than the AUC for baseline ($p = 0.01$) and post-therapy ($p = 0.03$) features alone. When only baseline features were used, the AUC was 0.64 and when only post-therapy features were used, the AUC was 0.60.

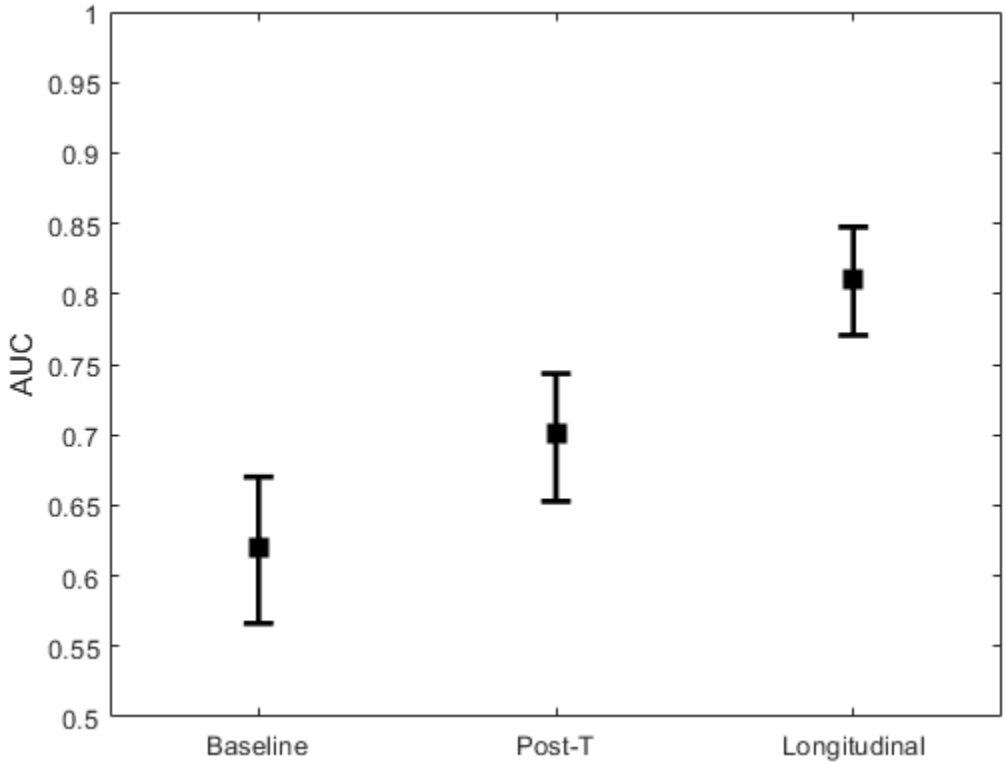


*Figure 40 – Comparison between the model performance using only baseline feature information, only post-therapy feature information, and longitudinal $iSUV_{max}$ feature information. The error bars represent the 95% confidence interval.*

#### *4.3.3.3 Lesion Classification Accuracy*

The lesion classification accuracy (Figure 41) using the trained RF model was superior to the accuracy of classifying lesions by any other response category approach based on a single feature (all $p < .003$).

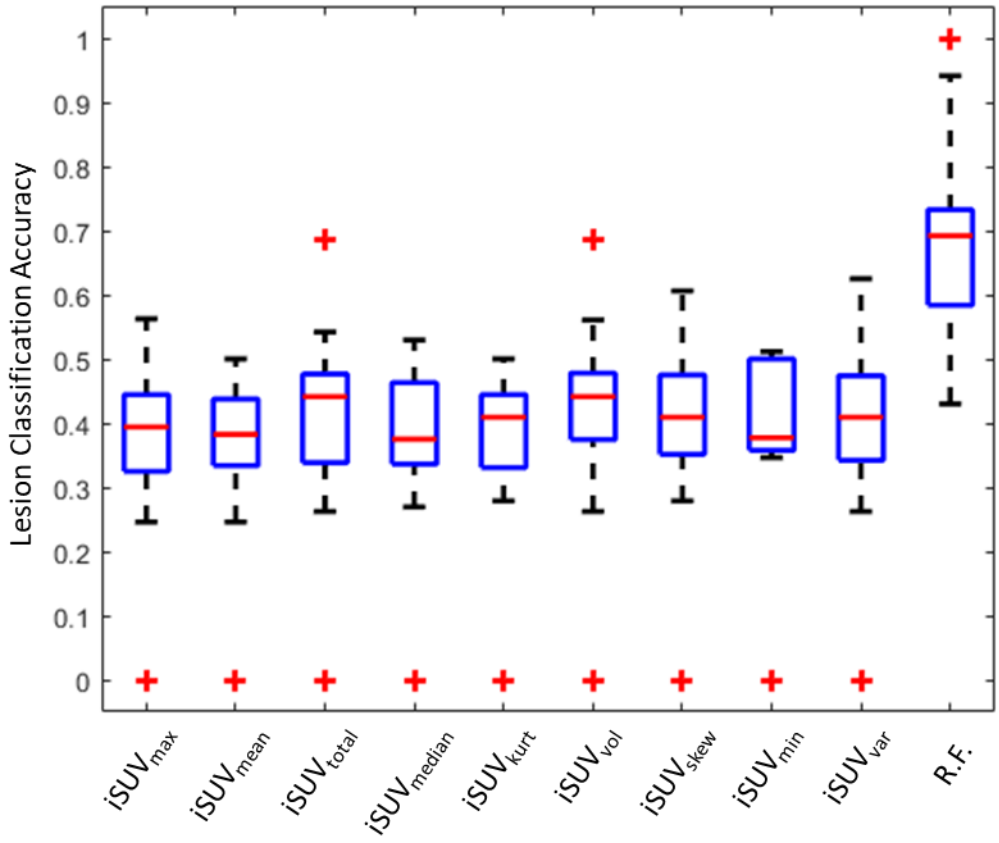


*Figure 41 – Lesion classification accuracy for different used methods. The single feature methods were based on early lesion response categories.*

### 4.3.4 Patient-Level Response Prediction

#### *4.3.4.1 Feature Selection and PFS Modeling*

The feature engineering process generated 8,162 unique, non-zero patient-level features, from which ten were selected according to the method of section 4.2.2.5. The feature number sensitivity analysis results are shown in Figure 42. The model performance initially increased with the number of features but plateaued at 6 features. Model prediction accuracy decreased starting from seven to ten features. The maximal C-Index of 0.842 occurred when both six and seven features were selected. However, the number of patients per selected feature must be also taken into consideration, a small number of patients per selected feature leads to overfitting of the model (Vittinghoff & McCulloch, 2007). Using $N_{opt} = 6$ results in six patients per selected feature, while using $N_{opt} = 4$ results in a much more reliable operating point of nine patients per selected

feature. The 2% (0.016) gain in accuracy (from 0.826 to 0.842) does not justify the increase in $N_{opt}$ and likely decrease in model robustness to other patient populations. Therefore, we have chosen $N_{opt} = 4$ features as the desirable operating point. The classification, survival analysis, sensitivity analyses, and the benchmarking were performed using a 4 feature MLR model.

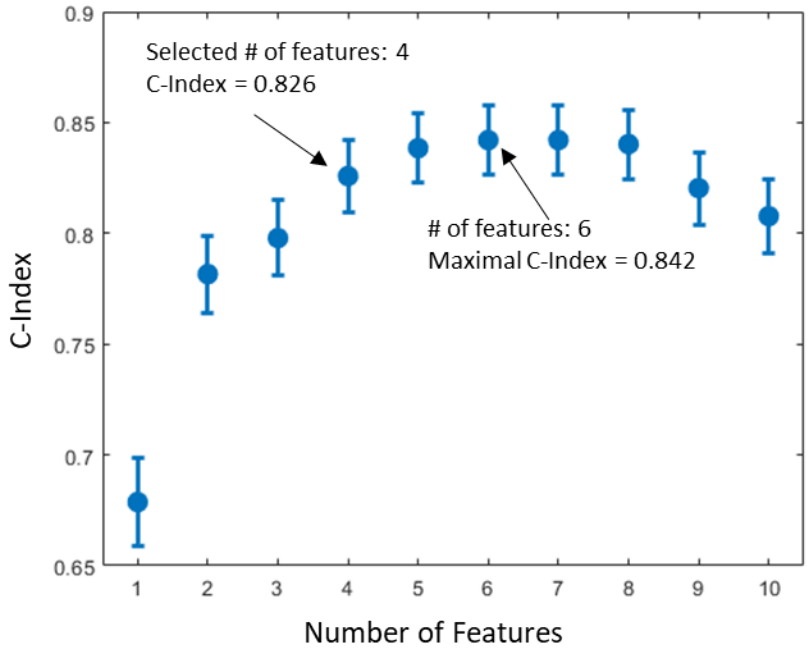


*Figure 42 – C-Index as a function of the number of features used to train the multivariate linear models. Each point in the plot corresponds to the accuracy of a model trained with the optimal set of its x-value number of features.*

The four features selected for the model are shown in Table 22. An *a posteriori* explanation of the clinical meaning of the selected features is presented in the discussion section. The correlation between the selected features was mild due to the minimum redundancy algorithm employed, the highest correlation in absolute values was found between F2 and F4 ($\rho = 0.4$). Table 23 shows the inter-correlation between the selected features. The model parameters estimated without cross-validation ($\hat{P}$) are reported to allow for the reproduction of the model.

*Table 22 – Features selected for the MLR modeling of PFS. The estimated parameters $\hat{P}$ did not consider leave-one-out cross validation and were not used for predictions within this work.*

| F# | Lesion-level feature | Norm. | Time | A.L. | P.S. | Patient Level Statistic | Correlation to PFS | $\hat{P}$ | Range |
|---|---|---|---|---|---|---|---|---|---|
| F1 | $iSUV_{mean}$ | Liver | Post-T | WB | NL | Min | -0.49 | -20.7 | [0, 1.2] |
| F2 | $iSUV_{var}$ | No | Variation | Spine | All | Median | -0.47 | -1.7 | [-6.8, 6.8] |
| F3 | $iSUV_{mean}$ | Spleen | Baseline | Chest | All | Min | 0.42 | 43.3 | [0, 0.7] |
| F4 | $iSUV_{var}$ | No | Post-T | Head | PL | Sum | -0.40 | -0.34 | [0, 50] |

Norm. = Normalization; A.L. = Anatomical Location; P.S. = Persistence Status; $\hat{P}$ = Estimated Parameters

*Table 23 - Correlation between the four selected features.*

| | F1 | F2 | F3 | F4 |
|---|---|---|---|---|
| F1 | 1 | 0.36 | 0.01 | 0.05 |
| F2 | | 1 | 0.06 | 0.40 |
| F3 | | | 1 | -0.07 |
| F4 | | | | 1 |

#### *4.3.4.2 Model Performance Assessment*

Using the selected features (Table 22), the predicted PFS ($\widehat{PFS}$) for each patient is shown in Figure 43. The fit of the model to the data showed a tendency of underestimating the PFS in the good responders' group and overestimating PFS in the poor responders' group (Figure 43C). The mean residual was 0.33 months with 12 months standard deviation.

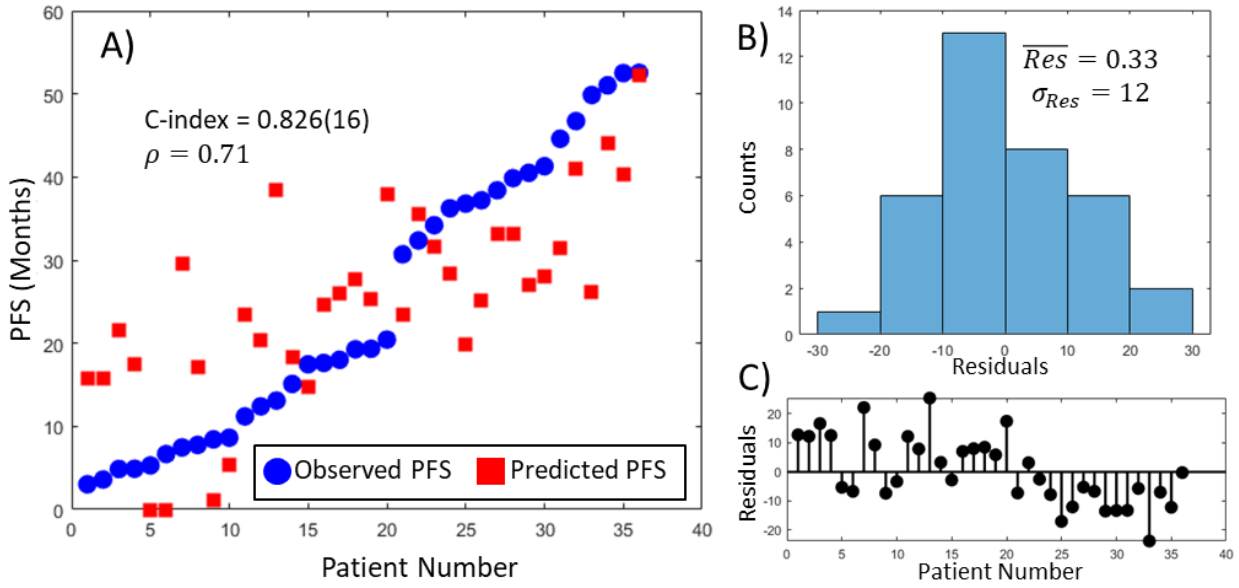


*Figure 43 – A) Results of PFS prediction compared to observed PFS, B) histogram of the prediction residuals with the mean and standard deviation, and C) absolute residuals per patient.*

The classification analysis resulted in an AUC of 0.88 with optimal operation point of 26 months, for which the specificity was 0.85 and sensitivity was 0.81 (Figure 44). Patient number 20 was removed from the classification analysis since this patient was the only one censored in the poor responder group and it is impossible to ascertain if this patient classification was correct.

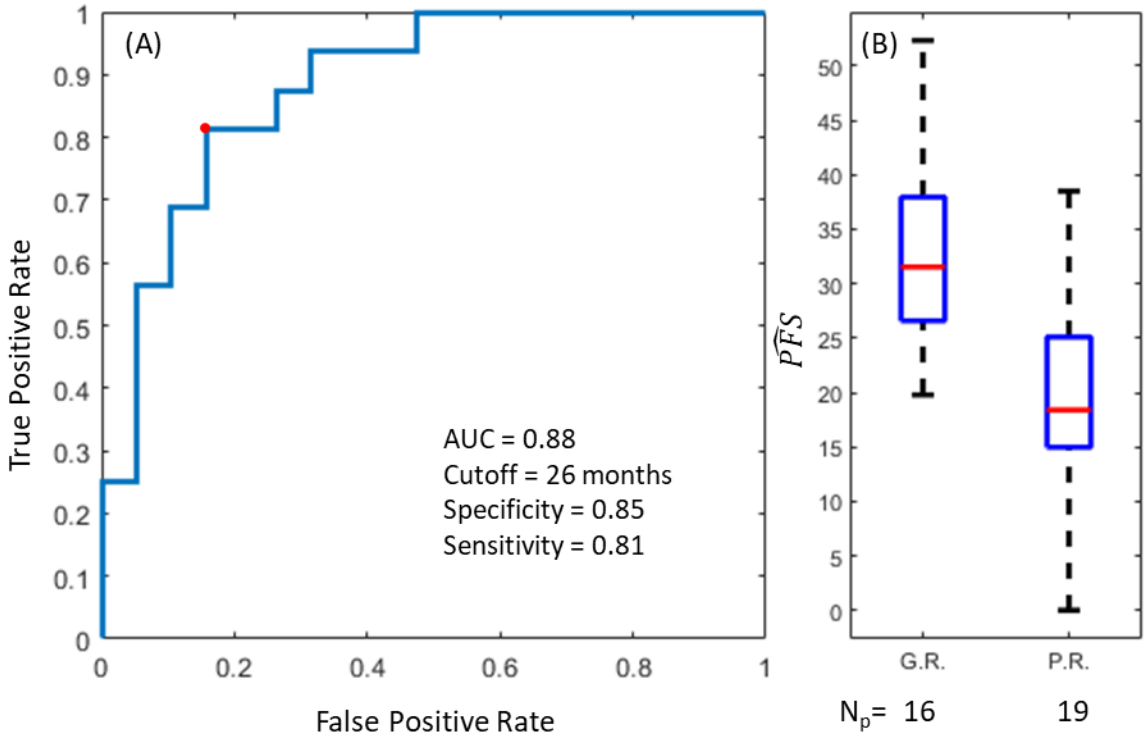


*Figure 44 – (A) Receiver operating curves for the classification of patients into good and poor responders based PFS prediction. The red dot indicates the optimal operating point. The area under the curve (AUC), specificity, and sensitivity at the optimal operating point are shown in the plot. (B) Shows the boxplot of the patients classified into good (G.R.) and poor responders (P.R.). Np= number of patients.*

The Kaplan-Meier plot is shown in Figure 45, the cutoff was 23 months, which resulted in a statistically significant difference between the groups ($p \ll .001$, log-rank test). The Cox proportional hazards ratio univariate analysis resulted in a normalized hazard ratio of $3\times 10^{-5}$ which was statistically significant ($p \ll .001$).

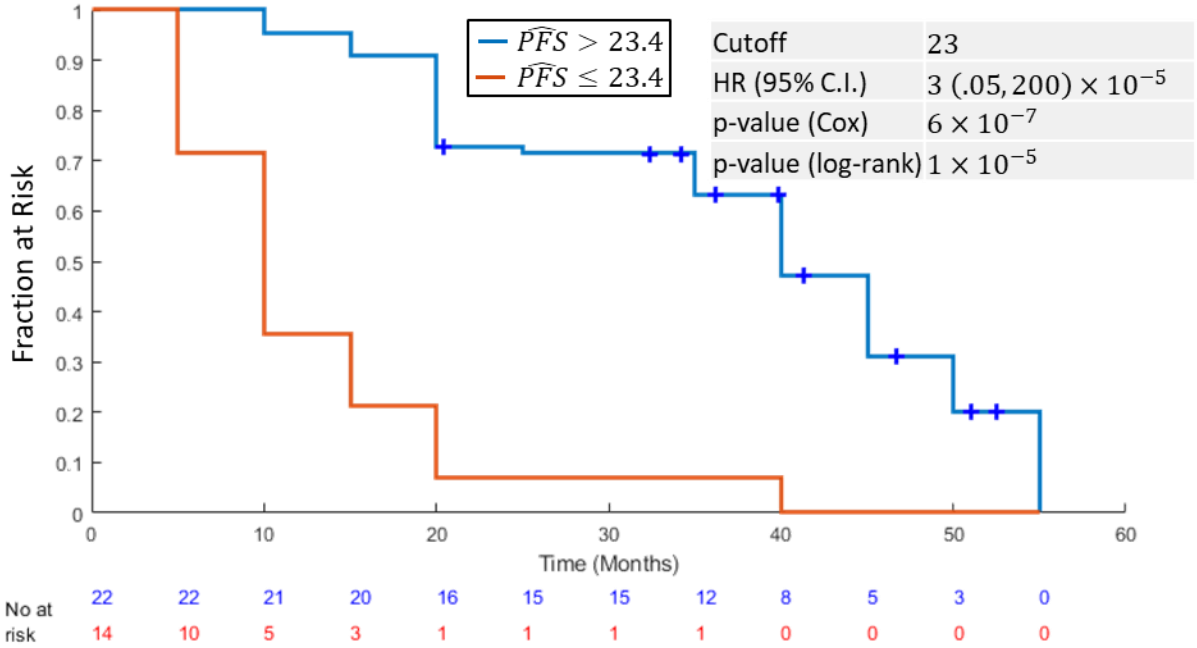


*Figure 45 - Kaplan-Meier curve using the MLR model predicted PFS as the group discriminator. The cutoff, hazard ratio (HR), p-value for a univariate Cox proportional hazards ratio model, and p-value for a log-rank test are shown on the plot.*

#### *4.3.4.3 Model Sensitivity to Selected Features*

Table 24 shows the alternative features that were used for each of the four features selected (Table 23). The aspects of each feature that vary from the original feature are highlighted in red.

*Table 24 - Features resulting from the feature variation degree analysis. The red highlights indicate aspects of each degraded feature that are different from the selected feature.*

| | Variation Degree | Lesion-level feature | Norm | Time | A.L. | P.S. | Patient level Statistic | $\rho_{sel}$ | $\rho_{PFS}$ |
|---|---|---|---|---|---|---|---|---|---|
| Feature #1 | Original feature | $SUV_{mean}$ | Liver | Post-T | WB | NL | Min | N/A | -0.49 |
| | 1 | $SUV_{max}$ | Liver | Post-T | WB | NL | Min | 0.94 | -0.44 |
| | 2 | $SUV_{max}$ | No | Post-T | WB | NL | Min | 0.89 | -0.48 |
| | 3 | $SUV_{mean}$ | No | Post-T | WB | NL | Min | 0.88 | -0.47 |
| | 4 | $SUV_{mean}$ | Spleen | Post-T | WB | NL | Min | 0.82 | -0.44 |
| Feature #2 | Original feature | $SUV_{var}$ | No | Variation | spine | All | Median | N/A | -0.47 |
| | 1 | $SUV_{var}$ | Liver | Variation | spine | All | Median | 0.94 | -0.43 |
| | 2 | $SUV_{max}$ | No | Variation | spine | All | Median | 0.74 | -0.19 |
| | 3 | $SUV_{max}$ | Liver | Variation | spine | All | Median | 0.72 | -0.21 |
| | 4 | $SUV_{max}$ | Liver | Variation | spine | All | Mean | 0.65 | -0.28 |
| Feature #3 | Original feature | $SUV_{mean}$ | Spleen | Baseline | Chest | All | Min | N/A | 0.42 |
| | 1 | $SUV_{mean}$ | No | Baseline | Chest | All | Min | 0.92 | 0.28 |
| | 2 | $SUV_{max}$ | Spleen | Baseline | Chest | All | Min | 0.91 | 0.31 |
| | 3 | $SUV_{mean}$ | Liver | Baseline | Chest | All | Min | 0.90 | 0.33 |
| | 4 | $SUV_{mean}$ | Spleen | Baseline | Chest | All | Mean | 0.83 | 0.40 |
| Feature #4 | Original feature | $SUV_{var}$ | No | Post-T | head | PL | Sum | N/A | -0.40 |
| | 1 | $SUV_{var}$ | No | Post-T | head | All | Sum | 0.97 | -0.39 |
| | 2 | $SUV_{var}$ | No | Post-T | head | PL | Max | 0.95 | -0.36 |
| | 3 | $SUV_{max}$ | No | Post-T | head | PL | Max | 0.94 | -0.40 |
| | 4 | $SUV_{var}$ | No | Post-T | head | All | Max | 0.93 | -0.34 |

Norm. = Normalization; A.L. = Anatomical Location; P.S. = Persistence Status; $\hat{P}$ = Estimated Parameters; $\rho_{sel}$ = correlation with the selected feature; $\rho_{PFS}$= correlation with the observed PFS; WB = whole-body

The correlation of the degraded features with the selected features, and that of the degraded features with the observed PFS (last two columns of Table 24) are shown in Figure 46. We can observe that each feature degrades at different rates and that the degraded features correlation with PFS does not degrade monotonically.

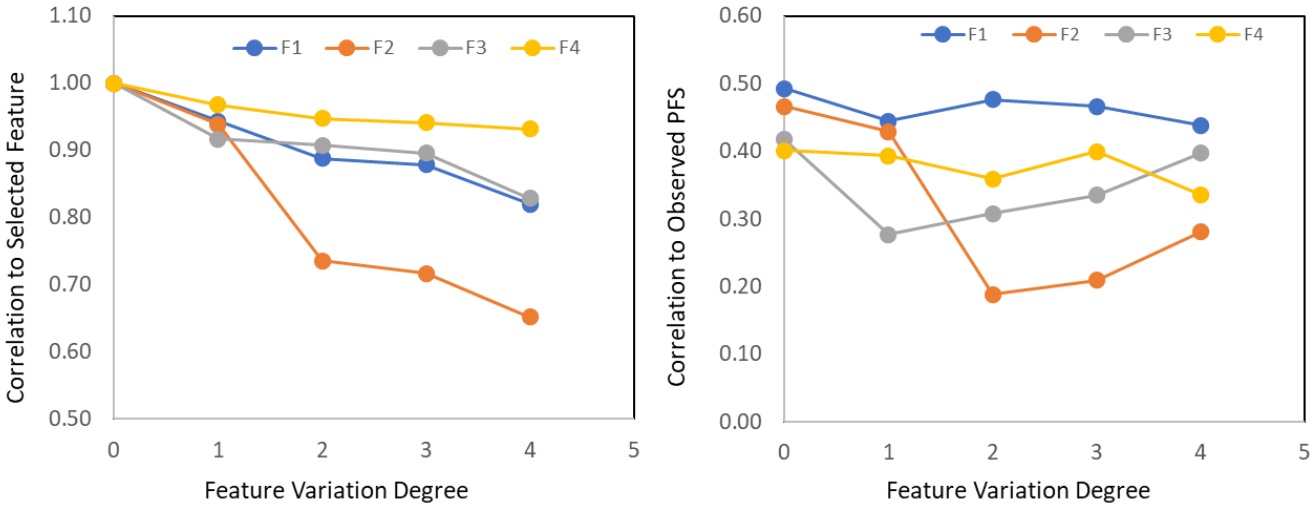


*Figure 46 - Plots of the correlations to the selected feature and to the observed PFS as a function of the feature variation degree.*

Finally, the feature selection sensitivity analysis results are shown in Figure 47. The median concordance decreases as the feature variation degree increases from 0 to 16. However, the degradation was mild, of only 0.081, from 0.826 at null feature variation degree to 0.745 at feature variation degree of 16. Additionally, there were C-Indices superior to 0.826, which is expected since the feature selection method did not optimize for model concordance but for feature correlation to $PFS$, another characteristic that prevents overfitting.

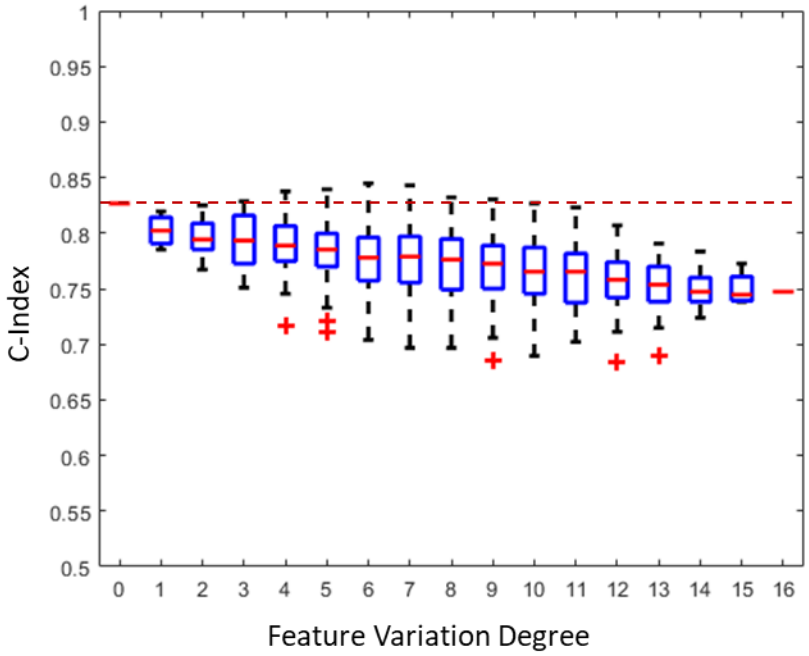


*Figure 47 – Boxplots of the C-Indices as a function of the feature variation degree. The dashed horizontal red line marks the C-Index = 0.826, which is the concordance index of the selected features.*

#### *4.3.4.4 Model Sensitivity to Image Analysis*

The results for the model sensitivity to image acquisition and analysis are summarized in Figure 48. The overall performance of the standard model was worsened by the removal of anatomical segmentation, lesion matching, lesion segmentation, and longitudinal imaging information (Figure 48). According to the reduction of the C-index, the most relevant aspect of image acquisition and analysis was the longitudinal information, without which a 16% reduction was observed in the C-index (from 0.826 to 0.691). The loss in concordance due to lack of longitudinal information was followed by lack of comprehensive lesion segmentation (-15%), lesion matching (-10%), and anatomical segmentation (-9%).

#### *4.3.4.5 Benchmarking*

To benchmark our results, we have conducted a literature review to identify all SRI features that were found to be significant in previous works. All considered SRI features and p-values from previous works are shown in Table 25.

*Table 25 – All considered features for benchmarking. Significant p-values are underlined*

| Features | Lesions # | Citation | Feat. # | Bonferroni α-value | Cox OS | Cox PFS | Log-rank OS | Log-rank PFS |
|---|---|---|---|---|---|---|---|---|
| $\Delta SUV_{TS}$ | 1 | (Haug et al., 2010) | 3 | 0.02 | | | | <u>0.002</u> |
| $SUV_{max}$ | 1 | (Campana et al., 2010) | 1 | 0.05 | | <u>0.001</u> | | <u>0.001</u> |
| $SUV_{max}$ | 1 | (Ambrosini et al., 2015) | 1 | 0.05 | | <u>0.003</u> | | <u>0.002</u> |
| $SUV_{max}$ | 5 | (Sharma et al., 2019) | 8 | 0.006 | | 0.03 | | |
| Entropy | 5* | (Werner et al., 2019) | 10 | 0.005 | 0.045 | | | |
| $SUV_{max}$ | all | (Carlsen et al., 2020) | 1 | 0.05 | | <u>0.005</u> | | |
| $SUV_{max}$ | all | (Carlsen et al., 2021) | 3 | 0.02 | 0.049 | | | |
| $\Delta mean(SUV_{max})$ | 5 | (Sharma et al., 2019) | 8 | 0.007 | | 0.024 | | |
| $SUV_{mean}$ | all | (E. Pauwels et al., 2020) | 10 | 0.005 | 0.024 | | | |
| $SUV_{mean}$ | all | (Ortega et al., 2021) | 36 | 0.0014 | | 0.0053 | | |
| $SUV_{mean}$ | all | (E. Pauwels et al., 2020) | 10 | 0.005 | 0.024 | | | |
| $SUV_{vol}$ | all | (E. Pauwels et al., 2020) | 10 | 0.005 | 0.037 | | | |
| $SUV_{vol}$ | all | (Carlsen et al., 2021) | 3 | 0.02 | <u>0.0001</u> | <u>0.001</u> | | |
| $SUV_{vol}$ | all | (Durmo et al., 2022) | 10 | 0.005 | <u>0.001</u> | | | |
| $min(SUV_{mean})$ | all | (Carlsen et al., 2021) | 3 | 0.02 | <u>0.001</u> | <u>0.001</u> | <u>0.001</u> | |
| $mean(SUV_{max})$ | 5 | (Ortega et al., 2021) | 36 | 0.0014 | | 0.023 | | |
| $mean(SUV_{TS})$ | 5 | (Ortega et al., 2021) | 36 | 0.0014 | | 0.028 | | |
| $mean(SUV_{TL})$ | 5 | (Ortega et al., 2021) | 36 | 0.0014 | | 0.047 | | |
| Skewness | all | (Ortega et al., 2021) | 36 | 0.0014 | | 0.017 | | |
| Kurtosis | all | (Ortega et al., 2021) | 36 | 0.0014 | | 0.022 | | |
| $\Delta SUV_{vol}$ | 5 | (Urso et al., 2023) | 18 | 0.003 | | 0.011 | | 0.006 |

OS = Overall Survival; PFS = Progression-free survival; α-value = Significance Value

We have used a Bonferroni corrected p-value, after which only some features were still considered significant. According to this significance analysis, the four features selected to be used as comparison to the model approach developed in this work are shown in Table 26.

*Table 26 – Features selected for benchmarking.*

| Features | # Lesions | Citation |
|---|---|---|
| $\Delta \mathrm{i}SUV_{T/S_{rel}}$ | 1 | (Haug et al., 2010) |
| $\max(iSUVmax)$ | all | (Campana et al., 2010)<br>(Ambrosini et al., 2015)<br>(Carlsen et al., 2020) |
| $sum(iSUVvol)$ | all | (Carlsen et al., 2021)<br>(Durmo et al., 2022)<br>(E. Pauwels et al., 2020) |
| $\min(iSUVmean)$ | all | (Carlsen et al., 2021) |

The PFS prediction approach introduced in this work outperformed all four single feature approaches tested even in the worst-case scenario (no longitudinal information). (Figure 48) None of the single feature approaches resulted in statistically significant AUC or Hazard ratios, whereas the standard MLR model approach classification was significant.

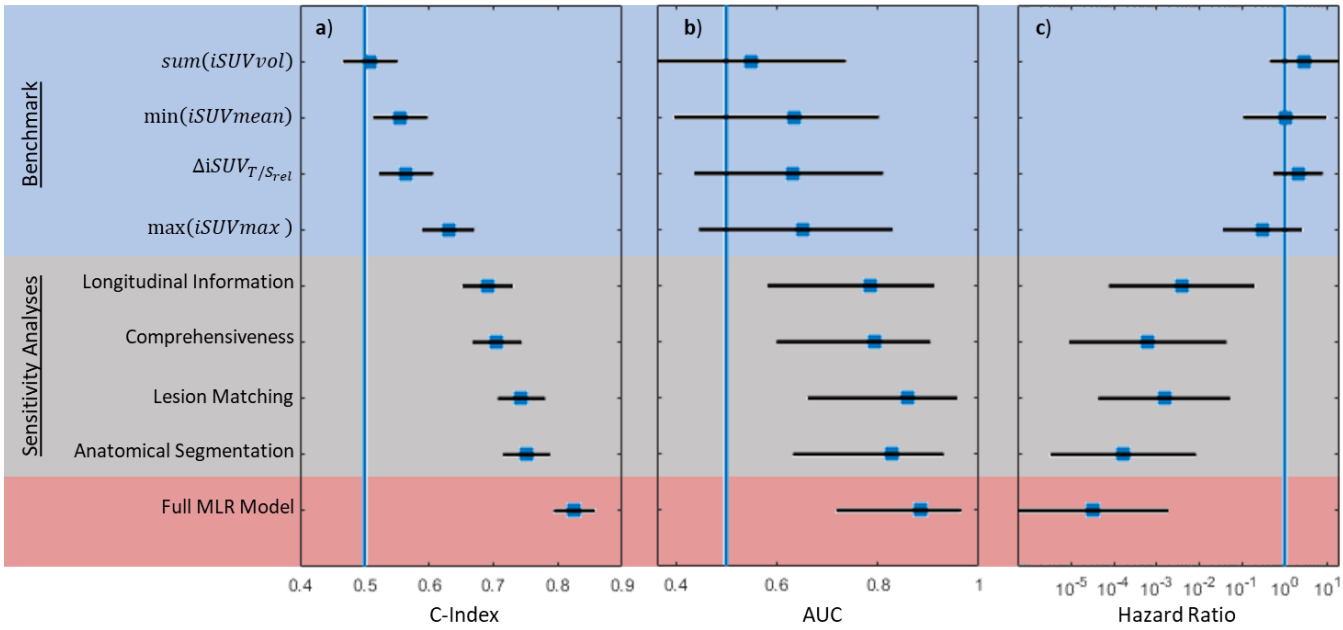


*Figure 48 –Benchmarking and sensitivity analyses results for univariate (A) C-Index, (B) AUC, and (C) Cox proportional hazards ratios. The black horizontal lines display the 95% confidence interval. The vertical blue line shows the value where the hazard ratio is equal to unity, and AUC and C-index are equal to 0.5, hence non-significant. The blue shaded area refers to the benchmark analysis results, the gray shaded area to the model sensitivity to image analysis results, and the red shaded area shows the full MLR model results.*

Table 27 shows the main quantitative results and the p-values for the statistical tests. Since we have tested 9 variables in this work, the Bonferroni corrected p-value is 0.006. This correction is conservative in that it neglects the correlation between the tested features.

*Table 27 – Compiled results for the Cox proportional hazards ratio analysis, the receiver operating curve analysis, and the Kaplan-Meier analysis. Significant p-values are underlined. The blue shaded area refers to the benchmark analysis results and the gray shaded area to the model sensitivity to image analysis results.*

| Feature | HR (95% C.I.) | p-value Cox | AUC | p-value AUC | AUC cutoff | p-value Log-Rank | Cutoff Log-Rank |
|---|---|---|---|---|---|---|---|
| $sum(iSUV_{vol})$ | 3 (0.4, 19) | .3 | .55 (.36, .73) | .8 | 240 | .4 | 117 |
| $\min(iSUV_{mean})$ | 1 (0.1, 10) | .99 | .63 (.40, .80) | .2 | 2.0 | .09 | 1 |
| $\Delta iSUV_{T/S_{rel}}$ | 2(0.5,8) | .3 | .63 (.44, .81) | .3 | 0.04 | .004 | 0.05 |
| $\max(iSUV_{max})$ | 0.3 (0.04,3) | .3 | .65 (.45, .83) | .2 | 46 | .009 | 37 |
| Longitudinal Information | $4(0.07,200)\times10^{-3}$ | <u>.006</u> | .78 (.58, .91) | .009 | 22 | <u>.0006</u> | 26 |
| Comprehensiveness | $6(0.09, 400)\times10^{-4}$ | <u>.0007</u> | .79 (.60, .90) | .002 | 24 | .01 | 29 |
| Lesion Matching | $1(0.04, 50)\times10^{-3}$ | <u>.0003</u> | .86 (.66, .96) | <u>.0001</u> | 22 | <u>.0001</u> | 22 |
| Anatomical Segmentation | $2(0.03, 80)\times10^{-4}$ | <u>.00001</u> | .83 (.63, .93) | .002 | 25 | <u>.0001</u> | 23 |
| Full MLR model | $3(0.05, 200)\times10^{-5}$ | <u>$6\times10^{-7}$</u> | .88 (.72, .97) | <u>$4\times10^{-5}$</u> | 26 | <u>$1\times10^{-5}$</u> | 23 |

## 4.4 Discussion

In this work we have used a comprehensive, lesion-level, longitudinal image analysis approach to perform lesion- and patient-level treatment response prediction. On the lesion-level, we have shown that developing simple models based on SRI PET features leads to accurate prediction of long-term persisting lesions (accuracy of 0.78 - Figure 38). We have also shown that the models are robust in relation to the PET feature used in the prediction (Figure 36), and that a model trained with one longitudinal feature pair performed as well as models that use multiple longitudinal feature pairs (Figure 35). In addition, our results show that the longitudinal information of early treatment response was vital for the prediction accuracy, with AUC falling from 0.82 to 0.70 (post-therapy only) and to 0.62 (baseline only) when longitudinal information was disregarded (Figure 40). Additionally, RF model-based lesion classification into persistent (PL) and disappearing lesions (DL) was more accurate than using response categories based on single SRI features (Figure 41). Which indicates that the lesions showing an increase in uptake despite having received PRRT are not necessarily likely to be persistent.

On the patient level, we have developed an MLR model predictive of the PFS of mNET patients receiving PRRT. Following a sensitivity analysis, four relevant features were selected (Figure 42, Table 23) in a compromise between the predictive performance and generalizability. The MLR model concordance index was 0.826 (Figure

43) and the AUC for classification of patients into poor and good responders was 0.88 with 0.85 specificity and 0.81 sensitivity (Figure 44). The predicted PFS significantly stratified the patient population into poor and good responders on a Kaplan-Meier Analysis (Figure 45, $p \ll .001$) and the hazard ratio of $3 \times 10^{-5}$ was also significant (Figure 45, $p \ll .001$). However, the residual standard deviation was high (12 months), showing that the model is accurate for patient stratification and classification but the precision for continuous prediction of PFS needs to be refined. The sensitivity analyses showed that our approach is robust to the selected features, with mild loss of concordance when correlated features were used to replace the selected features (Figure 47). Furthermore, the results worsening is mild (max decrease on 16%) due to variations in the image acquisition and analysis methods (Figure 48). While, the PFS prediction concordance and classification accuracy are impacted negatively by the removal of image analysis steps, it is not critical since they remain significant (Figure 48, Table 27). However, utilizing every step of the image acquisition and analysis yielded the best set of results. Lastly, in a benchmark study, the MLR modeling approach was shown to outperform the most promising single feature predictors (Table 26) in terms of classification and survival prediction (Figure 48, Table 27). From a modeling perspective, the $\widehat{PFS}$ is a prediction of the true patient PFS. However, from a clinical point of view, this is a prognostic rather than predictive biomarker since its purpose is not to decide between using PRRT in lieu of an alternative treatment. It can be applied, as demonstrated, to stratify poor and good responders among NET patients that receive PRRT.

The correlation plot of Figure 34 shows that many of the extracted SRI lesion-level features are strongly correlated. This agrees with two observations of this work: (1) that the lesion-level models are robust in relation to the longitudinal feature pair used in the prediction (Figure 36), and (2) that a lesion-level model trained with one longitudinal feature pair performs as well as models that use multiple longitudinal feature pairs (Figure 35). The strong feature correlation presents an explanation for why the addition of more features is unlikely to add relevant new information and thus leads to very similar results. On the lesion-level, we have elected $iSUV_{max}$ to be the feature of the optimal model for its simplicity of interpretation. The feature that showed the weakest

correlation with $iSUV_{max}$ is the $iSUV_{total}$. It might be expected that the model trained with these two longitudinal feature pairs would outperform the single feature model. We have conducted an *a posteriori* evaluation of such a model and found that an RF model with 10 splits using these two features achieved AUC of 0.81 (0.77, 0.84). Therefore, the performance achievable by using the two least correlated features is still not significantly better than the single feature performance (AUC of 0.78) and requires a more complex and likely less generalizable model.

In the patient-level prediction, a data-driven approach to selection of features was used, therefore there was no *a priori* hypothesis about what features would be predictors. However, each of the selected features has a clinical significance that justifies (*a posteriori*) their predictive power. F3 indicates that when the uptake at baseline of chest lesions is high, the PFS tends to be higher, which agrees with the well-established fact that a high uptake on baseline SRI indicates efficiency of the PRRT. Conversely, F1 indicates that elevated uptake at the post-therapy scan contributes to shorter PFS, also agreeing with the well-established clinical interpretation that an elevated post-therapy uptake is associated to disease progression. F2 indicates that when the variance in spine lesions increases, PFS tends to be shorter. This is in line with literature findings since higher genetic heterogeneity is correlated to more aggressive tumors (Lou et al., 2022) and the genetic characteristics of tumors can be captured by uptake heterogeneity on PET imaging (Basu et al., 2011). Furthermore, the negative impact of tumor heterogeneity into NET patients' outcome has been previously observed (Graf et al., 2020). Finally, F4 suggests that the presence of post-therapy lesions in the head is an important factor that worsens PFS outcome, which is plausible given the vital importance of this anatomic location. Furthermore, the robustness analysis shows that varying F4 characteristics such as lesion measure or the patient level statistic has a small effect on the correlation with PFS (Table 24, Figure 46), which further indicates that the important aspect of this feature is its anatomical location. Of note, liver and abdomen weren't included as anatomical locations of any selected feature or any of their correlated variations (Table 24). This indicates that, being in the most common site for metastases, the liver and the abdomen lesions are not the distinguishing factor for outcome prediction.

The chest region might have been selected for F3 because it is the anatomical location with the most elevated number of lesions after the abdominal region, thus carrying relevant discriminatory information about treatment response. About the anatomical location of F2, the spine is a sensitive organ in which metastases can greatly disrupt patients' quality of life which might explain its selection.

Despite the lack of a direct functional relationship of $iSUV_{max}$ variation to lesion-level response (Figure 41), the RF model trained is also interpretable, as seen in the simple decision tree from Figure 39. The simplicity of the developed predictive model is encouraging. Lesions with a low $iSUV_{max}$ (<1.3) on the post-therapy image are likely to disappear. This observation suggests that the PRRT effects on the tumors can continue even after the radioactive agent delivery period is over. Among the lesions with $iSUV_{max}$ higher than 1.3 on the post-therapy image, the lesions with equally high $iSUV_{max}$ on baseline are likely to persist, while the ones with a small uptake on baseline are likely to disappear. Of note, more complex models did not predict the lesion-level response with higher accuracy.

There are important technical difficulties that explain and justify the lack of comprehensive, lesion-level, and longitudinal SRI features in the previously published patient-level treatment response prediction studies. First, the segmentation of all lesions in a 3D medical image is challenging and models capable of such have only recently been made possible due to advent of convolutional neural-networks (Carlsen et al., 2022; Isensee et al., 2021; Wehrend et al., 2021). Furthermore, the use of lesion-level longitudinal information requires automated lesion-tracking since establishing manual correspondence between lesions is time consuming, prone to error and subjectivities (Santoro-Fernandes et al., 2021). Bridging these difficulties, together with the MLR modeling approach, are the main novelties of the present work.

Campana et al. and Durmo et al. have evaluated patient stratification based on SRI features (Campana et al., 2010; Durmo et al., 2022). In a retrospective study that enrolled 47 patients, Campana et al. observed an AUC of 0.75 (0.61, 0.9) for $max(iSUV_{max})$, while we observed an AUC of 0.65 (0.45, 0.83). In a retrospective study that enrolled 46 patients, Durmo et al. observed an AUC of 0.77 (0.62, 0.88) for $sum(iSUV_{vol})$, while we

observed an AUC of 0.55 (0.36, 0.73). In both studies the AUC was significant, while we found both to be non-significant when applied to this work's patient population. This puts in question the reproducibility of the single feature results. Importantly, fewer patients were used in the present work, however the order of magnitude of the patient populations was the same. This comparison is also limited by differences in the theranostics agents used, Durmo et al. used $^{68}$Ga-DOTATOC as an imaging agent and the PRRT was delivered with $^{177}$Lu- and $^{90}$Y-DOTATOC. While Campana et al. used the $^{68}$Ga-DOTANOC as an imaging agent and it is unclear what type of PRRT was used. It is possible that the features they identified are more suitable for the specific theranostics compounds used. However, the AUC obtained in this work using the $\widehat{PFS}$ was higher than the AUC obtained by Campana et al. and by Durmo et al, which begs the question if the modeling approach could be of value for other theranostics compounds.

In a recent publication, Laudicella et al. have evaluated the predictive power of two-dimensional SRI features to classify lesions into progressive vs non-progressive disease (Laudicella et al., 2022). In their definition, a progressive lesion had an increase of 25% in $iSUV_{max}$ or in area. They investigated logistic regression models to combine SRI features and predicted the NET lesion-level response to PRRT. However, the performance of the proposed models was inferior to that of baseline iSUVskew, which they identified as the best predictor with AUC of 0.74 for a cutoff of 2.45. Using our data, $iSUV_{skew}$ was not a good direct predictor of progressive vs. non-progressive lesion on long-term image, achieving an AUC of 0.57 (0.51, 0.62). Additionally, using $iSUV_{skew}$ as a direct predictor led to an accuracy of 0.45, specificity of 0.52, and sensitivity of 0.64. Furthermore, $iSUV_{skew}$ led to the smallest AUC result for a single feature RF model in our work (Figure 36). Therefore, our results suggest that the reproducibility of the predictive power observed by Laudicella et al. is limited, which might be explained by the large number of variables tested in their work for predictability of lesion-level response (65 variables).

Model robustness, reproducibility, and interpretability are fundamental aspects in the investigation of predictive biomarkers and of this work. During the feature extraction and the training of the models, we have

used methods that prevent overfitting and therefore promoted the robustness and reproducibility of the results, on the lesion-level model we have used 5-fold cross validation, an ensemble of 30 decision trees, and limiting the number of splits in each tree. On the patient-level models we have used 5-fold cross validation for feature selection and leave-one-out cross validation for the model fitting. We were also attentive to the number of features we tested to predict the outcomes and we corrected our significance value according to the number of tested features, which further corroborates to the reproducibility of our results. Despite these efforts, the predictive power of the features and models presented here must be evaluated on an independent patient population to be confirmed before any prospective clinical trial is suggested.

Some limitations arise from the retrospective nature of the present work. For one, the patients have confounding treatments and have not been monitored on a regular interval, which means that the observed PFS values are likely overestimated. Also, PFS was used instead of overall survival (OS) due to the difficulty of establishing OS in NET patients and because PFS is the most appropriate surrogate end point for NET patients (Strosberg et al., 2023). Furthermore, different scans and acquisition protocols were used, which introduces uncertainties to the measure SRI features. On top of that, the automated image analysis steps employed are also subject to uncertainties, namely the lesion segmentation, anatomical segmentation, and the lesion matching. The accuracies of the image analysis methods have been studied and reported elsewhere, lesion matching accuracy was studied in chapter 2, and lesion segmentation accuracies were reported by other authors (Schott et al., 2023; A. Weisman et al., 2022). Despite the existence of inaccuracies in each image analysis step, our results show that high prediction and classification performances can be obtained by utilizing the image information generated. This is indicative of the methodology robustness to acquisition protocols and to inaccuracies in the image analysis steps.

In this work we have assumed that a decrease in SRI uptake and the eventual disappearance of lesion conspicuity in this type of imaging modality is indicative of a favorable response to treatment. This assumption must be made carefully since the reduction in SRI uptake can mean a dedifferentiation of the tumor, which

would indicate an increase in cancer aggressiveness (Sundin & Rockall, 2012). The assumption that a decrease in SRI uptake is indicative of favorable response can be confirmed using anatomical imaging (e.g., contrast enhanced CT and MR) to investigate the same lesions. A conspicuous indication of cancerous tissue on an anatomical image, associated with a decrease in SRI uptake, consists of a strong suggestion of dedifferentiation and consequent increase in aggressiveness. Associating the anatomical and the functional response on a lesion-level is beyond the scope of the work presented herein and needs to be further investigated.

Another limitation was the variation present in the patient population regarding the number of PRRT cycles received per patient. Ten of the 36 patients received fewer than 4 cycles due to adverse events, death, or patient election to stop treatment. Evaluating the impact of the number of cycles in the patient outcome was beyond the scope of this work. However, this information is encoded in the longitudinal evaluation of SRI features, thus it is expected that the limitation in number of cycles will be reflected in the uptake information contained in the images. Therefore, it is reasonable to hypothesize that the developed model should not be less accurate for patients that received fewer PRRT cycles. Indeed, no statistically significant difference was observed in the model residuals between patients that received four vs. that received fewer than four PRRT cycles ($p = 0.49$, Wilcoxon rank sum test).

The dose-response correlations were indirectly used in this work to guide the prediction of patient-level response to treatment (PFS). The use of longitudinal feature variation from before and post-therapy images is likely to correlate to dosimetry estimated. Dose-response correlations for PRRT of NET patients have been observed, however, with widely variable responses for similar doses (Ilan et al., 2015; Jahn et al., 2020; S. Pauwels et al., 2005; Sgouros et al., 2021). At the same time, the correlation between the SRI uptake and delivered dose has also been observed, however limited (Bruvoll et al., 2023; Sainz-Esteban et al., 2012). Therefore, it is uncertain if using the estimated doses is likely to lead to more accurate outcome predictions than those obtained by longitudinal SRI features. Additionally, dosimetry is less feasible than the approach herein presented since dosimetry estimates require measurement of time-activity curves via time-series acquisitions

during the PRRT delivery, a process that is time-consuming, expensive, and can be uncomfortable to the patient due to prolonged clinical admission. Therefore, using the SRI as the source for extraction of response features is a more readily available and feasible approach to treatment response prediction than using the delivered dose estimate.

In this work we have explored SRI features only, features originating from other sources can be combined to SRI features and contribute to the value of prediction. For example, Pauwels et al. have explored the combination of the inflammation-based index (IBI) with $SUV_{mean}$ for its predictive power of PFS in PRRT of NET patients. Other non-imaging parameters have been shown to be extremely promising, notably the NETest, based on circulating tumor mRNA (Bodei et al., 2020). The MLR framework introduced in this work lends itself easily to the combination of predictive features originating from different sources. Furthermore, an alternative approach to the feature engineering used in this work would be to train regression models directly using the lesion graphs (section 3.2, Figure 20). Lesion graphs are a direct output of the image analysis used in this work and represent the entire information of disease course on a patient-level.

## 4.5 CONCLUSION

This is the first work to evaluate comprehensive, lesion-level, longitudinal SRI features three dimensionally, for all identifiable lesions of metastatic NET patients and to use these features to predict the lesion-level and patient-level response to PRRT. We have shown that by using simple models, the long-term persistence of lesions can be predicted with high accuracy. In the patient-level, our results show that comprehensive, lesion-level, longitudinal SRI analysis combined with multivariate statistical modeling leads to improved predictions of PRRT outcome in NET patients, significantly better than single-variate predictions. The suggested approach is promising, accurate, and robust. The lesion- and patient-level prediction capabilities can have important clinical implications in the decision making regarding systemic therapies and their combination with loco-regional treatments. However, we do not offer a final model for the prediction of lesion persistence or PFS in NET

patients receiving PRRT. To arrive at final models, a broader retrospective patient population must be analyzed, and a validation on untrained data must be performed, which will require multi-institutional collaborations. Finally, the clinical validation of such models will require prospective clinical trials.

# 5 GENERAL DISCUSSION

## 5.1 SUMMARY

The image-based treatment response assessment of metastatic cancer is currently limited to either a qualitative evaluation of the images or by a non-comprehensive quantification of lesion response. The overarching hypothesis of this work was that **comprehensive, lesion-level, and longitudinal imaging biomarkers will lead to improved treatment response prediction because of capturing disease response heterogeneity**. Therefore, the main goal of the presented thesis was **to develop an automated, comprehensive, lesion-level, longitudinal evaluation of metastatic cancer to enable enhanced treatment response assessment, and to evaluate the improvement in treatment response prediction both at the patient- and at the lesion-level.**

Image-based treatment response prediction, both on the patient- and the lesion-level, is normally performed with limited information due to a lack of methods for automated image analysis. In the work presented in this dissertation, we investigate the development and implementation of an automated method for longitudinal lesion tracking that extracts comprehensive, lesion-level, longitudinal information from medical images (chapters 2 and 3). This methodology was based on medical image registration and morphological operations to establish longitudinal correspondence of lesions identified at various time-points. Care was taken to assure a satisfactory performance in the setting of metastatic cancer longitudinal assessment. First, various cancer types were used to benchmark our methodology, mainly metastatic melanoma and neuroendocrine tumor, but also upper tract urothelial, rectal, breast, lung (non-small cell), prostate, colon, ovarian, appendiceal, adenoid, and head&neck (squamous cell). Second, lesion clustering was included to account for merging and splitting of lesions because of disease response to treatment. Third, the method accounted for multiple imaging modalities to allow for the utilization of medical images commonly acquired during routine patient care. We have established that deformable registration approaches lead to better lesion matching performance than rigid

approaches (chapter 2) and that an automated dilation strategy outperforms fixed dilations (chapter 3). The overall accuracy of lesion matching and tracking was always above 90%, and all sensitivity analyses evidenced the robustness of the method in various aspects. In one sensitivity analysis, we have shown that the lesion matching accuracy remained high for patients with high disease burden. In another one, we have shown that the accuracy remains constantly high across different anatomical location.

We then investigated the use of this information in treatment response prediction both on the lesion- and the patient-level (chapter 4). Comprehensive, lesion-level, and longitudinal image features, automatically obtained by the methodology described in chapters 2 and 3, were used to predict treatment response in mNET patients. A data-driven approach was used to extract SRI features from the PET/CT and PET/MR images of mNET patients. Features were engineered and selected, and lesion- and patient-level predictive models were built using the selected features.

Each aspect of the lesion matching methodology employed to obtain the SRI features was investigated regarding its importance in providing the model with predictive power. These included longitudinal image acquisition, individual lesion matching, anatomic lesion labeling, and complete lesion segmentation. Every step of information acquisition was shown to contribute to the final predictions performed by the model. Importantly, we have shown that the model's predictive power is robust in relation to the feature selection process, meaning that the information acquisition aspect of the work is crucial, and there can be variation as to what feature is selected, still maintaining the model predictive capability. This robustness is essential for clinical translation of the prediction models. The lesion-level predictions a long-term SRI time-point was used as the reference for lesion-level outcome and SRI features from the baseline and post-therapy time-points were used to fit simple random forest (RF) machine learning models. The resulting final RF model was relatively simple, interpretable (with a maximum of 4 splits per tree), and able to accurately predict the long-term persistence of a lesion (accuracy of 0.78). On the patient-level response prediction, a multivariate linear regression model was fitted, and its predictive power was benchmarked against single imaging feature prediction, an approach

previously investigated by other authors. The superiority of the proposed method was established (0.88 vs. 0.65 AUC for patient stratification), encouraging the further refinement of the models generated via this approach.

Our work presents several methodological innovations. Innovations for lesion matching include: the first use of deformable registration for lesion matching purposes; the first work using geometrical considerations, beyond lesion overlap, for integration of lesion merging and splitting to the lesion matching analysis; and the first work utilizing non-greedy optimization to enhance lesion matching assignment on whole-body images. Our technique is also innovative in the use of lesion-tracking across multiple time-points and across images of different modalities. Furthermore, this is the first use of comprehensive, longitudinal, and lesion-level metrics to derive predictive biomarkers of response in mNET patients receiving a targeted radionuclide therapy agent, and use them to predict treatment response in NETs by evaluating all of the lesions identified in a patient (Chapter 4). This is also the first use of three-dimensional molecular imaging information to predict long-term lesions persistence in widely spread mNET patients undergoing targeted radionuclide therapy (Chapter 4).

The methodology developed in this dissertation lays the foundation for a completely novel treatment response assessment framework. The work herein developed takes the treatment response assessment from a qualitative and subjective level to a quantitative and objective level. Importantly, the quantitative and objective information added by this methodology does not prevent qualitative input from experts. Instead, it serves as additional information upon which the experts can enhance their judgements of each patient case. It also takes the response assessment from a partial and incomplete level to a comprehensive level of analysis, which is the only way to fully characterize the heterogeneity in treatment response of metastatic solid cancers.

The key aspects of this methodology are that it is automated and that it is robust. First, the methodology has the potential to revolutionize the clinical treatment response assessment of metastatic cancer because of precluding extensive human intervention, which makes it suitable for addition to the clinical routine, most likely without burdening the clinical staff. However, the use of such information in clinical routine is evidently contingent on demonstration of clinical benefit. More importantly, we have shown that the methodology is

robust (chapters 2 and 3). In other words, the automated assessment fails rarely, and when it does, it fails non-catastrophically. This is a vital aspect for any automation tool that intends to be clinically translated. We have further shown that the non-perfect automated response assessment still leads to highly accurate predictions of patient- and lesion-level response to treatment (chapter 4). Predictions obtained with the introduced methodology outperformed predictions made with other non-comprehensive, or non-lesion-level, methodologies. Therefore, the research explored in this dissertation increases the extent to which medical images can serve the treatment response assessment and advances the patient and lesion-level response prediction framework to include comprehensive, lesion-level, and longitudinal information. However, the *de facto* use of such information for clinical purposes is evidently contingent on patient clinical benefit demonstration on prospective clinical trials.

## 5.2 Future work: Technology Development

### 5.2.1 Further The Lesion Tracking Development

Some methodological capabilities were not developed and could be explored in the future to enhance the usability of the developed image analysis tools. One such capability is the lesion tracking of images with intersecting but unequal field-of-views (FOV), which can be explored by adding an FOV localizer step before the initial rigid-registration. Another capability is tackling the issue of different arms positioning. This was not an impediment to the work presented herein because lesions in the arm region were not prevalent, as discussed in chapter 3. However, it did require the visual investigation of each case, which could be a hindrance for clinical translation. Strategies that identify arm positioning and perform separate registration approaches for the core and the arms volumes could be pursued to solve this issue. Finally, an important development to facilitate the clinical translation of the method is establishing the degree of certainty of the lesion matches and of the lesion tracks. This would serve to guide manual lesion track corrections, which could improve the degree of trust that healthcare professionals put on the tool for the detailed treatment response assessment and could speed up

manual intervention to the automated procedure. The manual correction of lesion tracks flagged as uncertain would also provide a solution for the observed trend of lower performance for long tracks, as shown in Figure 27 and discussed for the lesion tracks presented in Figure 28.

### 5.2.2 Lesion Graphs as Direct Predictors of Patient Outcome

To advance the use of longitudinal and lesion-level information in treatment response prediction, the utilization of lesion graphs as direct inputs for predictive models of patient outcomes such as OS and PFS should be explored. Lesion graphs have been proposed recently (Yan et al., 2018) and are in early stages of development (Kuckertz et al., 2022). The idea ties naturally with the lesion tracking concept. In a lesion graph, each layer represents a scan time-point, each node represents a lesion, and each edge represents a match. A schematic illustration of the concept is shown in Figure 49. Methodologies for using graphs as input for data-driven classification and prediction models have recently been developed (Cangea et al., 2018; Duvenaud et al., 2015; Hanik et al., 2022). The rationale for their use is that a lesion graph intrinsically encodes a great level of information relevant when evaluating a patient response: number of lesions per time-point, lesion response categories (responding, stable, progressing, new, disappearing), intensity of SUV uptake of each of the lesions in each time-point, and the complete dynamic of the longitudinal uptake progression. Furthermore, additional quantitative and qualitative information from the anatomic images can be included when available, e.g., anatomical location label, attenuation properties from the CT, or diffusion coefficient from diffusion weighted MR.

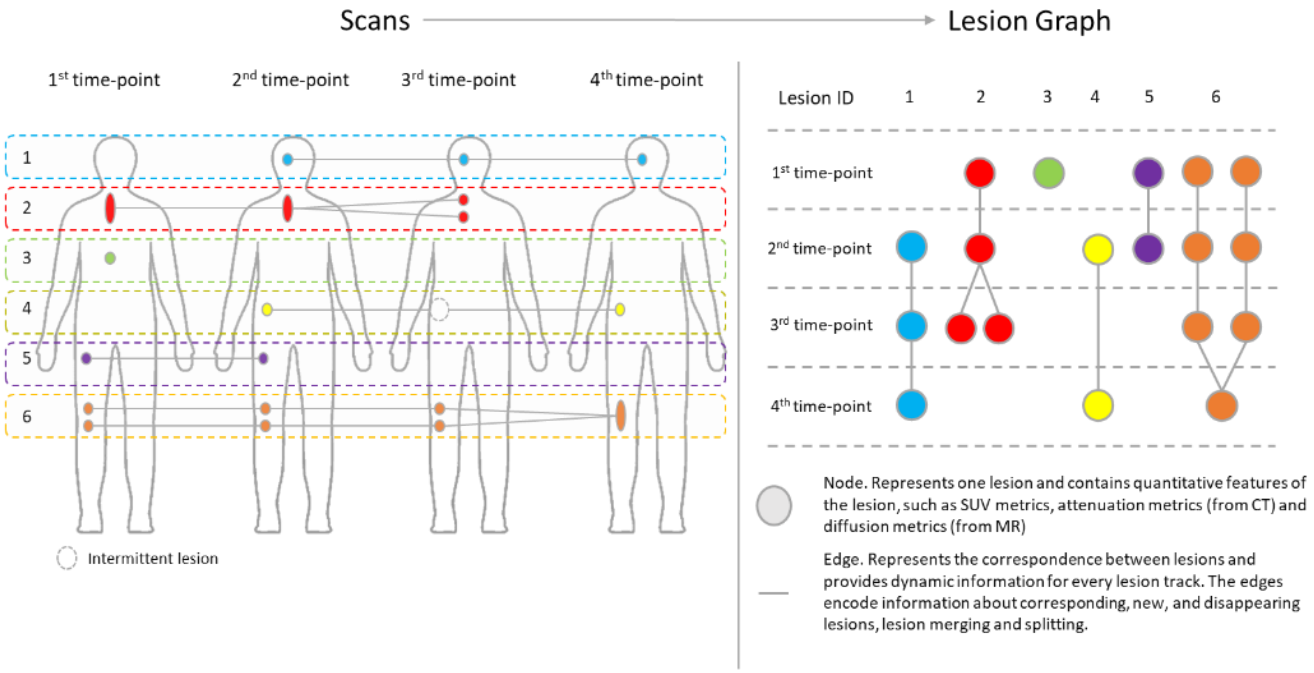


*Figure 49 – schematic representation of a lesion graph.*

This work's exploration of the lesion-tracking method was limited to a maximum of three time-points. The real power of the method shows when more time-points are evaluated. For example, in the Sankey plot of Figure 50. The Sankey plot shows the lesion response category evolution over-time and was made possible by the method introduced in chapters 2 and 3. The response categories indicated in the Sankey plot were, partial responder lesion (PRL), stable lesion (SL), and progressing lesion (PL). These categories were determined based on the change in $SUV_{total}$ from one scan to the next. A 30% change threshold was used like the description of section 4.2.2.6. The new lesions (NL) that appear through time are also shown in the Sankey plot, as well as the lesions that completely respond and therefore disappear (CRL). CRL, PRL, and SL were grouped as favorable responders and PL and NL as unfavorable. This type of treatment response visualization encompasses a greater amount of information than was previously possible without the lesion tracking method. It also further evidences the prominence of lesion-level response heterogeneity in mNET patients. It has great potential to influence clinical decision making about patient treatment management. These impacts need to be further evaluated with clinical trials and by building software that enables their clinical utilization by clinicians.

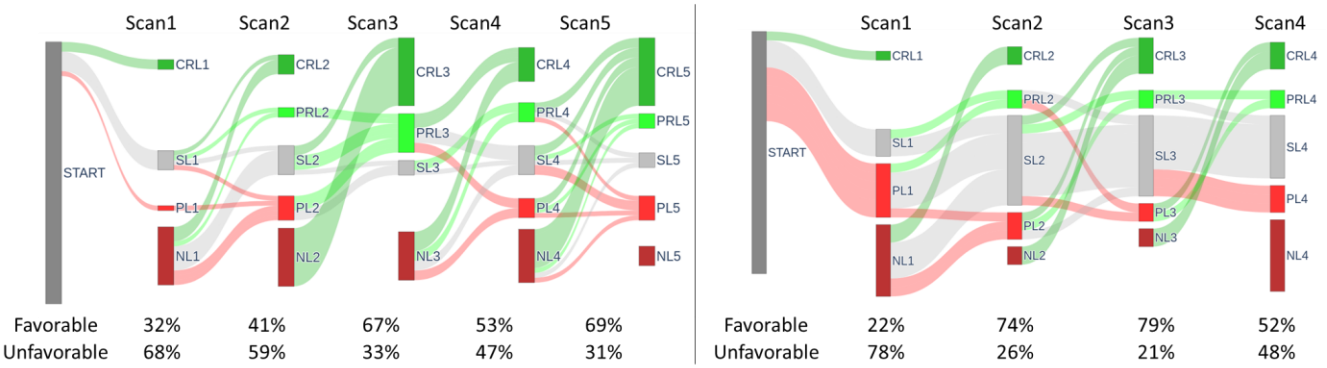


*Figure 50 – Sankey plots of two example mNET patients imaged five and four times with PET/CT $^{68}$Ga-DOTATATE over the course of PRRT. CRL = complete responder lesion, PRL = partial responder lesion, SL = stable lesion, PL = progressing lesion, NL = new lesion.*

### 5.2.3 Increasing the Models' Robustness

It is important to further the robustness study of the lesion-level and the patient-level treatment response prediction models developed in this work (chapter 4). To this end, its application in different mNET patient populations acquired at different institutions should be investigated. One way to pursue this goal is to establish international, multi-institutional collaborations to expand the models' applications to existing mNET patient populations (Figure 51, (Brady & Enderling, 2019)).

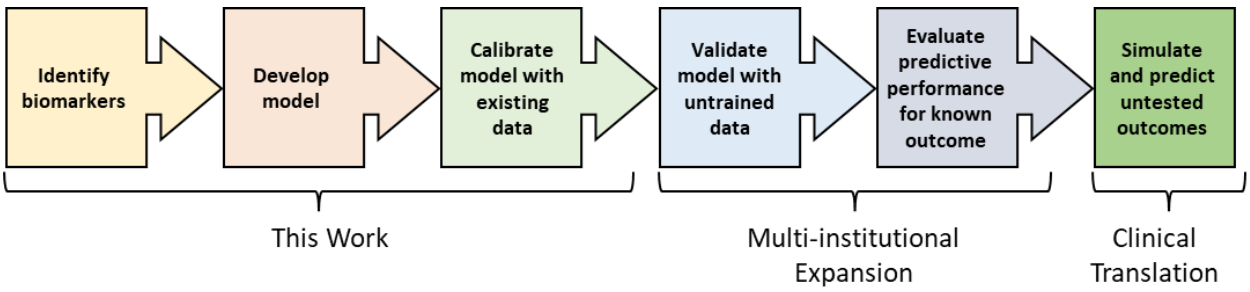


*Figure 51 – Path towards the clinical translation of cancer models, adapted from Brady & Enderling (Brady & Enderling, 2019)*

Table 28 shows results from a literature survey in which we assessed studies with goals and mNET patient-populations similar to the one presented in chapter 4. The patient-level MLR model herein developed could be further validated on these patient populations. In the survey, 264 patients were identified whose data could be

used to validate the predictive models. The possible methos validation on such a number of patients would greatly increase the level of confidence of the proposed models and, more importantly, it would allow independent cross-institutional validation of the models on untrained data, a crucial step that must precede the clinical model translation Figure 51. Because of its retrospective nature, these patient populations would have significant differences, the most significant ones being: institution specific scanners and acquisition protocols, image acquisition timings and agents, and the used PRRT agent. Strategies could be explored to factor these differences into the models or to evaluate the model robustness to these variations. As future steps of the development and validation of models predictive of patient benefit, we plan to take advantage of the Network for Imaging eXcellence (NIX) environment to establish research collaborations with these identified authors and institutions. NIX is a global initiative that emerged from Wisconsin Oncology Network of Imaging eXcellence (WONIX), to foster collaboration and data sharing amongst researchers that pursue the investigation and dissemination of quantitative imaging methodologies.

*Table 28 – mNET patient populations included in studies similar to this dissertation's*

| Citation | Baseline | Interim | post-T | Imaging Agent | PRRT Agent |
|---|---|---|---|---|---|
| (Gabriel et al., 2009) | 46 | 46 | | $^{68}$Ga-(TOC) | $^{90}$Y-(TOC) or $^{177}$Lu-(TATE) |
| (Haug et al., 2010) | 33 | 33 | | $^{68}$Ga-(TATE) | $^{90}$Y-(TATE), $^{177}$Lu-(TATE) |
| (Campana et al., 2010) | 47 | | | $^{68}$Ga-(NOC) | Unclear |
| (Öksüz et al., 2014) | 40 | | 40 | $^{68}$Ga-(TOC) | $^{90}$Y-(TOC) |
| (Ambrosini et al., 2015) | 43 | | | $^{68}$Ga-(NOC) | Unclear |
| (Kratochwil et al., 2015) | 30 | | | $^{68}$Ga-(TOC) | $^{90}$Y-(TATE), $^{177}$Lu-(TATE) |
| (Werner et al., 2017) | 141 | | | $^{68}$Ga-(TATE), $^{68}$Ga-(TOC) | $^{177}$Lu-(TATE), $^{177}$Lu-(TOC) |
| (Werner et al., 2019) | 31* | | | $^{68}$Ga-(TATE), $^{68}$Ga-(TOC) | $^{177}$Lu-(TATE) |
| (Sharma et al., 2019) | 55 | | 55 | $^{68}$Ga-(TATE) | $^{177}$Lu-(TATE) |
| (Graf et al., 2020) | 65 | | | $^{68}$Ga-(TATE), $^{68}$Ga-(TOC) | $^{177}$Lu-(TATE), $^{177}$Lu-(TOC) |
| (E. Pauwels et al., 2020) | 43 | | 43 | $^{68}$Ga-(TOC) | $^{90}$Y- (TOC) |
| (Ohlendorf et al., 2020) | 32 | 18 | | $^{68}$Ga-(TATE) | $^{177}$Lu-(TATE) |
| (Carlsen et al., 2020) | 128 | | | $^{64}$Cu-(TATE) | Unclear |
| (Carlsen et al., 2021) | 116* | | | $^{64}$Cu-(TATE) | Unclear |
| (Ortega et al., 2021) | 91 | 36 | | $^{68}$Ga-(TATE) | $^{177}$Lu-(TATE) |
| (Zwirtz et al., 2022) | 34 | | | $^{68}$Ga-(TATE), $^{68}$Ga-(TOC) | $^{177}$Lu-(TATE), $^{177}$Lu-(TOC) |
| (Opalińska et al., 2022) | 20 | | 12 | $^{68}$Ga-(TATE) | $^{177}$Lu-(TATE), $^{90}$Y-(TATE) |
| (Durmo et al., 2022) | 46 | 46 | | $^{68}$Ga-(TOC) | $^{90}$Y-(TOC), $^{177}$Lu-(TOC) |
| (Urso et al., 2023) | 39 | | 39 | $^{68}$Ga-(TOC) | $^{90}$Y-(TOC), $^{177}$Lu-(TOC) |
| (Laudicella et al., 2022) | 38 | | 38 | $^{68}$Ga-(TOC) | $^{177}$Lu-(TATOC) |
| U.W.C.C.C. | 109 | | 31 | $^{68}$Ga-(TATE) | $^{177}$Lu-(TATE) |
| R.P.C.C.C. | 30 | | 5 | $^{68}$Ga-(TATE) | $^{177}$Lu-(TATE) |
| | **1110** | **179** | **264** | | |

*Same patients used in a previous study by same authors, not included in the total. U.W.C.C.C. = university of Wisconsin Carbone Cancer Center; R.P.C.C.C. = Roswell Park Comprehensive Cancer Center

## 5.3 Future work: Possible Clinical Applications

### 5.3.1 Lesion-Level Prediction to Guide Loco-regional Treatment

The prediction of persistent lesions (chapter 4) has the potential to guide the delivery of subtotal locoregional therapy, possibly benefiting mNET patients. However, the extent of this possible benefit is unknown and needs to be investigated. The developed predictive model can be used to **simulate** the impact of selective subtotal locoregional therapy on reduction in liver tumor burden of mNET patients.

We have conducted an exploratory preliminary investigation of the potential benefit of introducing selective subtotal locoregional therapy to the PRRT systemic treatment framework for mNET patients. In this

exploratory study we simulated the ablation of persistent lesions and calculated the relative burden reduction on the long-term time-point it would cause. The simulated ablation of a given lesion consisted in assuming 100% reduction of the uptake (disappearance) of that lesion on the long-term image. The relative burden reduction was defined as the percent decrease of the total disease burden for a patient.

Preliminary results indicate that the use of the RF lesion persistence prediction leads to superior simulated burden reduction (Figure 52). The burden reduction for the RF model lesion selection was significantly higher than for all other single feature selections (all $p < .002$). These preliminary results encourage further investigation of lesion ablation simulation, which could possibly be used as a clinical decision support tool.

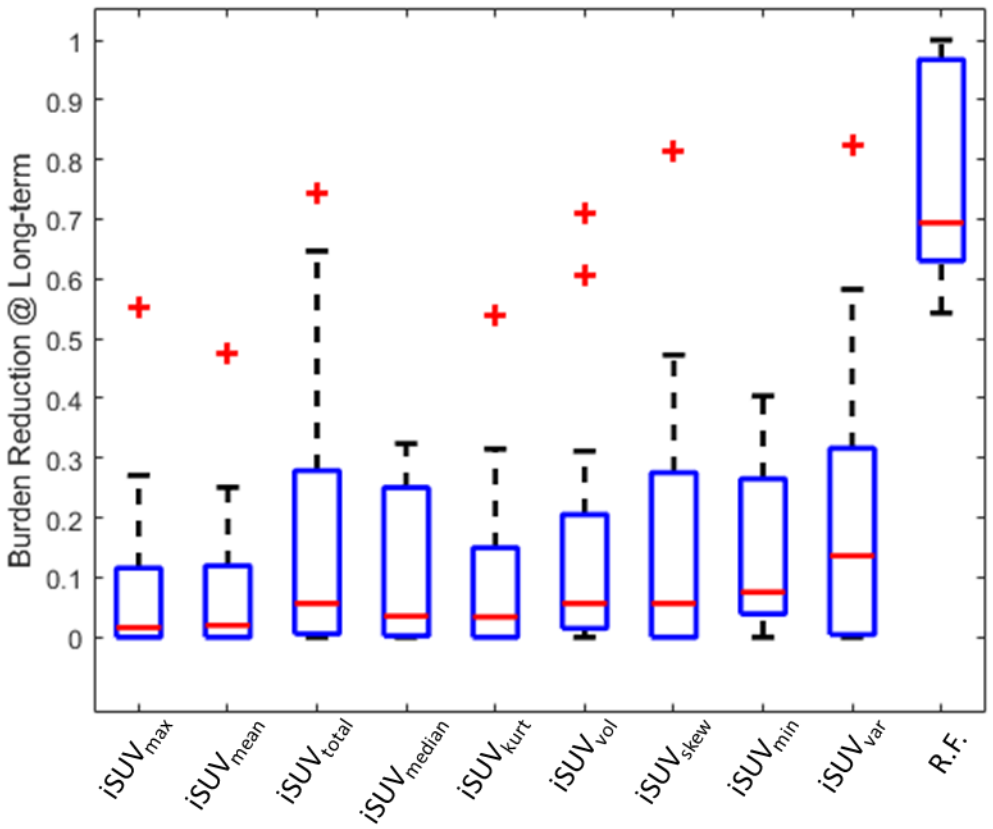


*Figure 52 – Simulation of the burden reduction caused by the selective ablation of the lesions according to single feature selection or to random forest (RF) model selection.*

Several improvements are necessary before the simulations of locoregional treatment can be considered for clinical support. Among them are: introducing evaluations of lesion-level feasibility of ablation, the ablation cost-benefit analysis, and tumor control probability evaluation. For the clinical feasibility study, interventional radiologists should be involved in the lesion selection process. One possible way of achieving that is that the radiologist would receive a visual ranking report indicative of the lesions that were selected as persistent. They would then decide which lesions can be ablated from a clinical perspective, considering factors dependent on intervention modality. For example, for stereotactic body radiation therapy (SBRT) these factors would be the number of fields, viability of treatment planning, and preservation of organs at risk. In the future, to assess the cost-benefit of lesion ablation, the estimated probability of lesion prevalence on long-term SRI derived from the quantitative image features (benefit – chapter 4) should be assessed against the feasibility and risk score provided by the physician (risk). To that end, a similar methodology to the one devised previously in our group for molecular imaging guided lesion biopsy could be explored (Harmon et al., 2016). Another aspect to consider is the expectation for lesion debulking and burden reduction. The hypothesis of 100% tumor burden reduction is simplistic. A probabilistic tumor control approach should be used in the simulations enrich the benefit models. For example, in liver metastases, the local control after SBRT ranges from 90% at 1 year to 76% after 3 years (Ohri et al., 2021). Incorporating this data into the models and simulations developed in this dissertation will make for a more accurate portrait of the reality of clinical intervention. Another factor to consider in the number of lesions considered for ablation. It is clear that ablating an elevated number of lesions leads to higher levels of morbidity (Elias et al., 2009). The final number of feasible lesions for ablation will likely depend on the intervention method chosen and must consider physician judgement, that can be supported by the review of the quantitative automated imaging reports.

### 5.3.2 Lesion-Level can Influence the Patient-Level Predictions

In chapter 4 of this dissertation, we used the comprehensive, lesion-level, longitudinal features to predict patient outcome in the form of PFS and the long-term lesion response of each individual lesion. It is reasonable

to hypothesize that the gain of information provided by the lesion-level response models could be factored into the patient-level models for an improved predictive performance. For example, lesion-level assessment allows for the investigation of relative importance for clinical outcome of metastases in different organs. Also, the lesion-level prediction of an elevated number of persisting lesions could be a driver for shorter prediction of PFS on the patient-level. Furthermore, the likely consequence of adding locoregional therapies could be estimated via the lesion-level models and then transposed to overall patient-level benefit. This approach to treatment benefit modeling would help patients and physicians to make better informed decisions regarding cancer treatment, weighing risks against the expected benefits. For the estimated impact in patient benefit to be accurate, it is necessary to improve and validate the models by increasing their specificity (for each possible locoregional therapy) and by conducting prospective interventional trials.

### 5.3.3 Applications Beyond mNET

The developed and validated approach to image analysis, feature engineering, and feature selection is not restricted in application to mNET patients. It can be applied to patients with other cancer types and receiving different systemic treatments. For each case, the application of the extracted features will depend on specific factors such as treatment adaptation possibilities, treatment, and image acquisition. Consequently, the investigation of each cancer type must be specific. The methodology can be used mostly for solid and metastatic cancers. For non-solid cancers, such as leukemias, the methodology would not be beneficial since spatial heterogeneity in response cannot be defined. Furthermore, for solid cancers that are commonly non-metastatic, such as glioblastomas, the methodology would be of limited benefit since its main usefulness is tied to the metastatic nature of the disease.

One example of a cancer type, beyond mNET, for which this methodology would be useful is metastatic melanoma. In fact, the methodology's development was initiated using a metastatic melanoma patient population (Chapter 2). Metastatic melanoma is the most lethal form of cutaneous neoplasm and its incidence continues to rise; there were more than 300,000 cases worldwide of MM in 2020 representing almost 2% of all

cancers (Switzer et al., 2022). It originates in the epithelial tissue, but the metastatic lesions can spread out through the entire patient anatomy, they are most commonly found in soft tissues but can also occur in bones. Chemotherapy with Dacarbazine was the standard of care in 2012 and had a response rate of 20% (Sun & Schuchter, 2001). That changed with the advent of immunotherapies which have a response rate of over 50% (Michielin et al., 2020). Immunotherapies of advanced metastatic melanoma patients have been introduced relatively recently (Bhatia et al., 2009) and the state-of-the-art immunotherapy for metastatic melanoma is ipi-nivo, the combination of an anti-PD1 (nivolumab) and an anti-CTLA-4 (ipilimumab) agent (Owonikoko et al., 2019). The treatment is administered regularly throughout periods that can go from a few months to a few years. Before, during, and after treatment, these patients are normally imaged with whole-body $^{18}$F-FDG PET/CT. However, the patient-level variation in response to immunotherapies is well-known but poorly understood (Saenger & Wolchok, 2008), prompting a need for improved response assessment technologies (Ralli et al., 2020). A recent doctoral dissertation has shown that the comprehensive, lesion-level, longitudinal functional medical image features can be used to predict treatment outcome and toxicities in this setting (Huff, 2022). In his doctoral work, Dr. Huff extracted the imaging features manually, a tedious and error-prone task that the approach introduced in this work can expedite.

Another example of a solid cancer whose study would highly benefit from the methodology developed in this dissertation is metastatic castrate resistant prostate cancer (mCRPC). Recently, response heterogeneity for metastatic prostate cancer has been observed to be correlated with patient response (Kyriakopoulos et al., 2020). In their work, Kyriakopoulos used a tracer ($^{18}$F-NaF) and a lesion matching tool that are limited to bone anatomy (Yip et al., 2014). In its most advanced stages, prostate cancer is likely to extensively metastasize to bone tissue, but soft tissue lesions are also present. Recently, the development of a targeted molecule using the prostate specific membrane antigen (PSMA) introduced a tracer that targets lesions both in bone and in soft tissue. Furthermore, it enabled advanced metastatic prostate cancer to be treated with the theranostics approach via the combination of $^{68}$Ga-PSMA for imaging and $^{177}$Lu-PSMA (approved in the U.S. by the FDA in

March of 2022) or $^{225}$Ac-PSMA (still not approved) for therapeutic targeted irradiation of the cancer cells (Hofman et al., 2018). This promising approach is similar to the one used with somatostatin receptors for mNET (PRRT) and therefore could equally benefit from comprehensive, lesion-level treatment response assessment. Methodologies for comprehensive identification and contouring of lesions already exist (Seifert et al., 2020) and the value of medical imaging in predicting treatment outcome has already been observed (Seifert et al., 2021). A comprehensive lesion-level analysis of the imaging information is a natural next step in the investigation of mCRPC and, as shown in this work, has the potential to increase the treatment predictability. Importantly, targeting resistant lesions with selective ablation is a research aim in a recently awarded grant from the specialized program of research excellence (SPORE) of the National Cancer Institute (NCI) and adaptations of the methods developed in this dissertation will be employed in this investigation.

In summary, the main disease characteristics that make the herein described methodology applicable and useful for disease response assessment are, (1) solid metastatic cancer with (2) poorly understood and heterogenous patient-level response to treatment, (3) also presenting lesion-level response heterogeneity, and (4) with various treatment options and possible treatment modification and adaptations. We do not mean to suggest that the adaptation of the analysis described herein would be straightforward, but that it is possible and desirable. The extension of the comprehensive, lesion-level, longitudinal image response assessment methodology to other cancers for which it is applicable would require detailed investigation of the usefulness of the methodology. Therefore, this dissertation work consists of the beginning of the investigations and developments into the novel, quantitative, and comprehensive manners of treatment response assessment and consequently of better-informed patient management.

# 6 CONCLUSION

The main contribution of this work was to establish a framework that allows for a more complete utilization of the information that is inherently present in medical images but currently under-utilized. This impacts cancer patients in multiple ways. For one, **improving the predictive power of treatment response** can enable oncologists to switch treatments early on if a satisfactory response is deemed unlikely, contributing to more personalized treatments. For another, **fully accounting for lesion-level heterogeneity in response to systemic treatments** enables better informed patient management. For example, identifying lesions that are candidates for selective ablation can provide the oncologist with tools to address non-responding lesions while maintaining the patient on a systemic treatment that is causing most lesions to be stable or respond. Lastly, **offering more practical surrogate endpoints for clinical trials** by monitoring each lesion individually can provide faster radiological endpoints to be used in clinical trials of novel drugs and vaccines, increasing efficiency, and reducing costs in the advancement of oncologic therapies. Furthermore, the developed methodology has great translational potential, and patient-specific decisions could be made possible upon this translation.

We aimed to investigate if an approach to extract comprehensive, lesion-level, longitudinal features was feasible and if it would improve the treatment response prediction. We have shown the feasibility of the information extraction by developing and validating a methodology that performs it, and we have shown the usefulness of the information generated by showing its ability to improve the prediction of patient and lesion-level outcome in cancer treatment in a metastatic setting. It has been suggested multiple times in the literature referenced in this dissertation that the comprehensive, lesion-level, longitudinal approach is the future of metastatic cancer treatment response assessment. Here, we have established the foundations to make it a clinical reality. By enabling the investigation of metastatic cancer on a comprehensive, lesion-level, and longitudinal manner, we are meaningfully contributing towards prolonged and higher-quality life of metastatic cancer patients.

# REFERENCES


Agarwal, M., Brahmanday, G., Bajaj, S. K., Ravikrishnan, K. P., & Wong, C. Y. O. (2010). Revisiting the prognostic value of preoperative 18F-fluoro-2- deoxyglucose (18F-FDG) positron emission tomography (PET) in early-stage (I & II) non-small cell lung cancers (NSCLC). *European Journal of Nuclear Medicine and Molecular Imaging*, *37*(4), 691–698. https://doi.org/10.1007/s00259-009-1291-x

Ahmed, M. (2020). Gastrointestinal neuroendocrine tumors in 2020. *World Journal of Gastrointestinal Oncology*, *12*(8), 791–807. https://doi.org/10.4251/wjgo.v12.i8.791

Akbarzadeh, A., Gutierrez, D., Baskin, A., Ay, M. R., Ahmadian, A., Riahi Alam, N., Lövblad, K. O., & Zaidi, H. (2013). Evaluation of whole-body mr to ct deformable image registration. *Journal of Applied Clinical Medical Physics*, *14*(4), 238–253. https://doi.org/10.1120/jacmp.v14i4.4163

Akyildiz, H. Y., Mitchell, J., Milas, M., Siperstein, A., & Berber, E. (2010). Laparoscopic radiofrequency thermal ablation of neuroendocrine hepatic metastases: Long-term follow-up. *Surgery*, *148*(6), 1288–1293. https://doi.org/10.1016/j.surg.2010.09.014

Al-Taan, O. S., Eltweri, A., Sharpe, D., Rodgers, P. M., Ubhi, S. S., & Bowrey, D. J. (2014). Prognostic value of Baseline FDG Uptake on PET-CT in Esophageal Carcinoma. *World Journal of Gastrointestinal Oncology*, *6*(5), 139–144. https://doi.org/10.4251/wjgo.v6.i5.139

Ambrosini, V., Campana, D., Polverari, G., Peterle, C., Diodato, S., Ricci, C., Allegri, V., Casadei, R., Tomassetti, P., & Fanti, S. (2015). Prognostic value of 68Ga-DOTANOC PET/CT SUVmax in patients with neuroendocrine tumors of the pancreas. *Journal of Nuclear Medicine*, *56*(12), 1843–1848. https://doi.org/10.2967/jnumed.115.162719

Apolo, A. B., Lindenberg, L., Shih, J. H., Mena, E., Kim, J. W., Park, J. C., Alikhani, A., McKinney, Y. Y., Weaver, J., Turkbey, B., Parnes, H. L., Wood, L. V., Madan, R. A., Gulley, J. L., Dahut, W. L., Kurdziel, K. A., & Choyke, P. L. (2016). Prospective study evaluating Na18F PET/CT in predicting clinical outcomes and survival in advanced prostate cancer. *Journal of Nuclear Medicine*, *57*(6), 886–892. https://doi.org/10.2967/jnumed.115.166512

Atkinson, C., Ganeshan, B., Endozo, R., Wan, S., Aldridge, M. D., Groves, A. M., Bomanji, J. B., & Gaze, M. N. (2021). Radiomics-Based Texture Analysis of 68Ga-DOTATATE Positron Emission Tomography and Computed Tomography Images as a Prognostic Biomarker in Adults With Neuroendocrine Cancers Treated With 177Lu-DOTATATE. *Frontiers in Oncology*, *11*(August), 1–9. https://doi.org/10.3389/fonc.2021.686235

Avants, B. B., Epstein, C. L., Grossman, M., & Gee, J. C. (2008). Symmetric diffeomorphic image registration with cross-correlation: Evaluating automated labeling of elderly and neurodegenerative brain. *Medical Image Analysis*, *12*(1), 26–41. https://doi.org/10.1016/j.media.2007.06.004

Basu, S., Kwee, T. C., Gatenby, R., Saboury, B., Torigian, D. A., & Alavi, A. (2011). Evolving role of molecular imaging with PET in detecting and characterizing heterogeneity of cancer tissue at the primary and metastatic sites, a plausible explanation for failed attempts to cure malignant disorders. *European Journal of Nuclear Medicine and Molecular Imaging*, *38*(6), 987–991. https://doi.org/10.1007/s00259-011-1787-z

Baum, R. P., Prasad, V., Müller, D., Schuchardt, C., Orlova, A., Wennborg, A., Tolmachev, V., & Feldwisch, J. (2010). Molecular Imaging of HER2 -Expressing Malignant Tumors in Breast Cancer Patients Using Synthetic 111 In- or 68 Ga-Labeled Affibody Molecules. *Journal of Nuclear Medicine*, *51*(6), 892–897. https://doi.org/10.2967/jnumed.109.073239

Becx, M. N., Minczeles, N. S., Brabander, T., de Herder, W. W., Nonnekens, J., & Hofland, J. (2022). A Clinical Guide to Peptide Receptor Radionuclide Therapy with 177Lu-DOTATATE in Neuroendocrine Tumor

Patients. *Cancers*, *14*(23). https://doi.org/10.3390/cancers14235792

Bhatia, S., Tykodi, S. S., & Thompson, J. A. (2009). Treatment of metastatic melanoma: An overview. *Oncology*, *23*(6), 488–496.

Bodei, L., Kidd, M. S., Singh, A., van der Zwan, W. A., Severi, S., Drozdov, I. A., Malczewska, A., Baum, R. P., Kwekkeboom, D. J., Paganelli, G., Krenning, E. P., & Modlin, I. M. (2020). PRRT neuroendocrine tumor response monitored using circulating transcript analysis: the NETest. *European Journal of Nuclear Medicine and Molecular Imaging*, *47*(4), 895–906. https://doi.org/10.1007/s00259-019-04601-3

Bosc, M., Heitz, F., Armspach, J. P., Namer, I., Gounot, D., & Rumbach, L. (2003). Automatic change detection in multimodal serial MRI: Application to multiple sclerosis lesion evolution. *NeuroImage*, *20*(2), 643–656. https://doi.org/10.1016/S1053-8119(03)00406-3

Brady, R., & Enderling, H. (2019). Mathematical Models of Cancer: When to Predict Novel Therapies, and When Not to. *Bulletin of Mathematical Biology*, *81*(10), 3722–3731. https://doi.org/10.1007/s11538-019-00640-x

Bruvoll, R., Blakkisrud, J., Mikalsen, L. T., Connelly, J., & Stokke, C. (2023). Correlations between [68Ga]Ga-DOTA-TOC Uptake and Absorbed Dose from [177Lu]Lu-DOTA-TATE. *Cancers*, *15*(4), 1–13. https://doi.org/10.3390/cancers15041134

Campana, D., Ambrosini, V., Pezzilli, R., Fanti, S., Labate, A. M. M., Santini, D., Ceccarelli, C., Nori, F., Franchi, R., Corinaldesi, R., & Tomassetti, P. (2010). Standardized uptake values of68Ga-DOTANOC PET: A promising prognostic tool in neuroendocrine tumors. *Journal of Nuclear Medicine*, *51*(3), 353–359. https://doi.org/10.2967/jnumed.109.066662

Cangea, C., Veličković, P., Jovanović, N., Kipf, T., & Liò, P. (2018). *Towards Sparse Hierarchical Graph Classifiers*. 1–6. http://arxiv.org/abs/1811.01287

Carlino, M. S., Saunders, C. A. B., Haydu, L. E., Menzies, A. M., Martin Curtis, C., Lebowitz, P. F., Kefford, R. F., & Long, G. V. (2013). 18F-labelled fluorodeoxyglucose-positron emission tomography (FDG-PET) heterogeneity of response is prognostic in dabrafenib treated BRAF mutant metastatic melanoma. *European Journal of Cancer*, *49*(2), 395–402. https://doi.org/10.1016/j.ejca.2012.08.018

Carlsen, E. A., Johnbeck, C. B., Binderup, T., Loft, M., Pfeifer, A., Mortensen, J., Oturai, P., Loft, A., Berthelsen, A. K., Langer, S. W., Knigge, U., & Kjaer, A. (2020). 64Cu-DOTATATE PET/CT and prediction of overall and progression-free survival in patients with neuroendocrine neoplasms. *Journal of Nuclear Medicine*, *61*(10), 1491–1497. https://doi.org/10.2967/jnumed.119.240143

Carlsen, E. A., Johnbeck, C. B., Loft, M., Pfeifer, A., Oturai, P., Langer, S. W., Knigge, U., Ladefoged, C. N., & Kjaer, A. (2021). Semi-automatic tumor delineation for evaluation of 64 Cu-DOTATATE PET/CT in patients with neuroendocrine neoplasms: prognostication based on lowest lesion uptake and total tumor volume. *Journal of Nuclear Medicine*, *62*(11), jnumed.120.258392. https://doi.org/10.2967/jnumed.120.258392

Carlsen, E. A., Lindholm, K., Hindsholm, A., Gæde, M., Ladefoged, C. N., Loft, M., Johnbeck, C. B., Langer, S. W., Oturai, P., Knigge, U., Kjaer, A., & Andersen, F. L. (2022). A convolutional neural network for total tumor segmentation in [64Cu]Cu-DOTATATE PET/CT of patients with neuroendocrine neoplasms. *EJNMMI Research*, *12*(1). https://doi.org/10.1186/s13550-022-00901-2

Cerfolio, R. J., Bryant, A. S., Winokur, T. S., Ohja, B., & Bartolucci, A. A. (2004). Repeat FDG-PET after neoadjuvant therapy is a predictor of pathologic response in patients with non-small cell lung cancer. *Annals of Thoracic Surgery*, *78*(6), 1903–1909. https://doi.org/10.1016/j.athoracsur.2004.06.102

Chaffer, C. L., & Weinberg, R. A. (2011). A Perspective on Cancer Cell Metastasis. *Science*, *331*(6024), 1559–1564. https://doi.org/10.1126/science.1203543

Cives, M., & Strosberg, J. R. (2018). Gastroenteropancreatic Neuroendocrine Tumors. *CA: A Cancer Journal for Clinicians*, *68*(6), 471–487. https://doi.org/10.3322/caac.21493

Crusz, S. M., Tang, Y. Z., Sarker, S. J., Prevoo, W., Kiyani, I., Beltran, L., Peters, J., Sahdev, A., Bex, A., Powles, T., & Gerlinger, M. (2016). Heterogeneous response and progression patterns reveal phenotypic heterogeneity of tyrosine kinase inhibitor response in metastatic renal cell carcinoma. *BMC Medicine*, *14*(1), 1–9. https://doi.org/10.1186/s12916-016-0729-9

Curran, S. D., Muellner, A. U., & Schwartz, L. H. (2006). Imaging response assessment in oncology. *Cancer Imaging : The Official Publication of the International Cancer Imaging Society*, *6*(October), 126–130. https://doi.org/10.1102/1470-7330.2006.9039

Dagogo-Jack, I., & Shaw, A. T. (2018). Tumour heterogeneity and resistance to cancer therapies. *Nature Reviews Clinical Oncology*, *15*(2), 81–94. https://doi.org/10.1038/nrclinonc.2017.166

De Mestier, L., Dromain, C., D'Assignies, G., Scoazec, J. Y., Lassau, N., Lebtahi, R., Brixi, H., Mitry, E., Guimbaud, R., Courbon, F., D'Herbomez, M., & Cadiot, G. (2014). Evaluating digestive neuroendocrine tumor progression and therapeutic responses in the era of targeted therapies: State of the art. *Endocrine-Related Cancer*, *21*(3). https://doi.org/10.1530/ERC-13-0365

Dong, Z.-Y., Zhai, H.-R., Hou, Q.-Y., Su, J., Liu, S.-Y., Yan, H.-H., Li, Y.-S., Chen, Z.-Y., Zhong, W.-Z., & Wu, Y.-L. (2017). Mixed Responses to Systemic Therapy Revealed Potential Genetic Heterogeneity and Poor Survival in Patients with Non-Small Cell Lung Cancer. *The Oncologist*, *22*(1), 61–69. https://doi.org/10.1634/theoncologist.2016-0150

Duffy, A. G., Ulahannan, S. V., Makorova-Rusher, O., Rahma, O., Wedemeyer, H., Pratt, D., Davis, J. L., Hughes, M. S., Heller, T., ElGindi, M., Uppala, A., Korangy, F., Kleiner, D. E., Figg, W. D., Venzon, D., Steinberg, S. M., Venkatesan, A. M., Krishnasamy, V., Abi-Jaoudeh, N., … Greten, T. F. (2017). Tremelimumab in combination with ablation in patients with advanced hepatocellular carcinoma. *Journal of Hepatology*, *66*(3), 545–551. https://doi.org/10.1016/j.jhep.2016.10.029

Durmo, R., Filice, A., Fioroni, F., Cervati, V., Finocchiaro, D., Coruzzi, C., Besutti, G., Fanello, S., Frasoldati, A., & Versari, A. (2022). Predictive and Prognostic Role of Pre-Therapy and Interim 68Ga-DOTATOC PET/CT Parameters in Metastatic Advanced Neuroendocrine Tumor Patients Treated with PRRT. *Cancers*, *14*(3). https://doi.org/10.3390/cancers14030592

Duvenaud, D., Maclaurin, D., Aguilera-Iparraguirre, J., Gómez-Bombarelli, R., Hirzel, T., Aspuru-Guzik, A., & Adams, R. P. (2015). Convolutional Networks on Graphs for Learning Molecular Fingerprints. *Advances in Neural Information Processing Systems*, *2015-Janua*, 2224–2232. http://arxiv.org/abs/1509.09292

Eisenhauer, E. A., Therasse, P., Bogaerts, J., Schwartz, L. H., Sargent, D., Ford, R., Dancey, J., Arbuck, S., Gwyther, S., Mooney, M., Rubinstein, L., Shankar, L., Dodd, L., Kaplan, R., Lacombe, D., & Verweij, J. (2009). New response evaluation criteria in solid tumours: Revised RECIST guideline (version 1.1). *European Journal of Cancer*, *45*(2), 228–247. https://doi.org/10.1016/j.ejca.2008.10.026

Elias, D., Goéré, D., Leroux, G., Dromain, C., Leboulleux, S., de Baere, T., Ducreux, M., & Baudin, E. (2009). Combined liver surgery and RFA for patients with gastroenteropancreatic endocrine tumors presenting with more than 15 metastases to the liver. *European Journal of Surgical Oncology*, *35*(10), 1092–1097. https://doi.org/10.1016/j.ejso.2009.02.017

Figiel, J. H., Viniol, S. G., Görlach, J., Rinke, A., Librizzi, D., & Mahnken, A. H. (2020). Update Regarding Imaging of Neuroendocrine Neoplasms. *RöFo - Fortschritte Auf Dem Gebiet Der Röntgenstrahlen Und Der Bildgebenden Verfahren*, *192*(02), 171–182. https://doi.org/10.1055/a-1001-2412

Fletcher, J. W., Djulbegovic, B., Soares, H. P., Siegel, B. A., Lowe, V. J., Lyman, G. H., Coleman, R. E., Wahl, R., Paschold, J. C., Avril, N., Einhorn, L. H., Suh, W. W., Samson, D., Delbeke, D., Gorman, M., & Shields, A. F. (2008). Recommendations on the use of 18F-FDG PET in oncology. *Journal of Nuclear Medicine*, *49*(3), 480–508. https://doi.org/10.2967/jnumed.107.047787

Freeborough, P. A., Woods, R. P., & Fox, N. C. (1996). Accurate registration of serial 3D MR brain images and its

application to visualizing change in neurodegenerative disorders. *Journal of Computer Assisted Tomography*, *20*(6), 1012–1022. https://doi.org/10.1097/00004728-199611000-00030

Gabriel, M., Oberauer, A., Dobrozemsky, G., Decristoforo, C., Putzer, D., Kendler, D., Uprimny, C., Kovacs, P., Bale, R., & Virgolini, I. J. (2009). 68Ga-DOTA-Tyr3-octreotide PET for assessing response to somatostatin-receptor-mediated radionuclide therapy. *Journal of Nuclear Medicine*, *50*(9), 1427–1434. https://doi.org/10.2967/jnumed.108.053421

Gallamini, A., Zwarthoed, C., & Borra, A. (2014). Positron emission tomography (PET) in oncology. *Cancers*, *6*(4), 1821–1889. https://doi.org/10.3390/cancers6041821

Gayed, I., Vu, T., Iyer, R., Johnson, M., Macapinlac, H., Swanston, N., & Podoloff, D. (2004). The role of 18F-FDG PET in staging and early prediction of response to therapy of recurrent gastrointestinal stromal tumors. *Journal of Nuclear Medicine*, *45*(1), 17–21.

Gerig, G., Welti, D., Guttmann, C. R. G., Colchester, A. C. F., & Székely, G. (2000). Exploring the discrimination power of the time domain for segmentation and characterization of active lesions in serial MR data. *Medical Image Analysis*, *4*(1), 31–42. https://doi.org/10.1016/S1361-8415(00)00005-0

Graf, J., Pape, U. F., Jann, H., Denecke, T., Arsenic, R., Brenner, W., Pavel, M., & Prasad, V. (2020). Prognostic Significance of Somatostatin Receptor Heterogeneity in Progressive Neuroendocrine Tumor Treated with Lu-177 DOTATOC or Lu-177 DOTATATE. *European Journal of Nuclear Medicine and Molecular Imaging*, *47*(4), 881–894. https://doi.org/10.1007/s00259-019-04439-9

H.H., C., G.J., C., K.W., K., J.W., K., N.-H., P., & Y.S., S. (2014). Preoperative PET/CT FDG standardized uptake value of pelvic lymph nodes as a significant prognostic factor in patients with uterine cervical cancer. *European Journal of Nuclear Medicine and Molecular Imaging*, *41*(4), 674–681. http://www.embase.com/search/results?subaction=viewrecord&from=export&id=L52867664%0Ahttp://dx.doi.org/10.1007/s00259-013-2626-1

Hafez, N., & Gettinger, S. (2020). Oligometastatic Disease and Local Therapies: A Medical Oncology Perspective. *Cancer Journal (United States)*, *26*(2), 144–148. https://doi.org/10.1097/PPO.0000000000000439

Hajnal, J. V., Saeed, N., Oatridge, A., Williams, E. J., Young, I. R., & Bydder, G. M. (1995). Detection of subtle brain changes using sub voxel registration and subtraction of serial mr images. In *Journal of Computer Assisted Tomography* (Vol. 19, Issue 5, pp. 677–691). https://doi.org/10.1097/00004728-199509000-00001

Hanik, M., Demirtaş, M. A., Gharsallaoui, M. A., & Rekik, I. (2022). Predicting cognitive scores with graph neural networks through sample selection learning. *Brain Imaging and Behavior*, *16*(3), 1123–1138. https://doi.org/10.1007/s11682-021-00585-7

Harmon, S. A., Tuite, M. J., & Jeraj, R. (2016). Molecular image-directed biopsies: Improving clinical biopsy selection in patients with multiple tumors. *Physics in Medicine and Biology*, *61*(20), 7282–7299. https://doi.org/10.1088/0031-9155/61/20/7282

Haug, A. R., Auernhammer, C. J., Wängler, B., Schmidt, G. P., Uebleis, C., Göke, B., Cumming, P., Bartenstein, P., Tiling, R., & Hacker, M. (2010). 68Ga-DOTATATE PET/CT for the early prediction of response to somatostatin receptor-mediated radionuclide therapy in patients with well-differentiated neuroendocrine tumors. *Journal of Nuclear Medicine*, *51*(9), 1349–1356. https://doi.org/10.2967/jnumed.110.075002

Hellier, P., Barillot, C., Corouge, I., Gibaud, B., Le Goualher, G., Collins, L., Evans, A., Malandain, G., & Ayache, N. (2001). Retrospective Evaluation of Inter-subject Brain Registration. In *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)* (Vol. 2208, Issue 9, pp. 258–265). https://doi.org/10.1007/3-540-45468-3_31

Hendlisz, A., Deleporte, A., Delaunoit, T., Maréchal, R., Peeters, M., Holbrechts, S., Van Den Eynde, M., Houbiers, G., Filleul, B., Van Laethem, J. L., Ceyssens, S., Barbuto, A. M., Lhommel, R., Demolin, G., Garcia, C., El Mansy, H., Ameye, L., Moreau, M., Guiot, T., … Santini, D. (2015). The prognostic significance of metabolic

response heterogeneity in metastatic colorectal cancer. *PLoS ONE*, *10*(9), 1–14. https://doi.org/10.1371/journal.pone.0138341

Hering, A., Peisen, F., Amaral, T., Gatidis, S., Eigentler, T., Othman, A., & Moltz, J. (2021). Whole-Body Soft-Tissue Lesion Tracking and Segmentation in Longitudinal CT Imaging Studies. *Proceedings of Machine Learning Research*, *143*, 312–326.

Hofman, M. S., Violet, J., Hicks, R. J., Ferdinandus, J., Ping Thang, S., Akhurst, T., Iravani, A., Kong, G., Ravi Kumar, A., Murphy, D. G., Eu, P., Jackson, P., Scalzo, M., Williams, S. G., & Sandhu, S. (2018). [ 177 Lu]-PSMA-617 radionuclide treatment in patients with metastatic castration-resistant prostate cancer (LuPSMA trial): a single-centre, single-arm, phase 2 study. *The Lancet Oncology*, *19*(6), 825–833. https://doi.org/10.1016/S1470-2045(18)30198-0

Hribernik, N., Huff, D. T., Studen, A., Zevnik, K., Klaneček, Ž., Emamekhoo, H., Škalic, K., Jeraj, R., & Reberšek, M. (2022). Quantitative imaging biomarkers of immune-related adverse events in immune-checkpoint blockade-treated metastatic melanoma patients : a pilot study. *European Journal of Nuclear Medicine and Molecular Imaging*, 1857–1869. https://doi.org/10.1007/s00259-021-05650-3

Huff, D. T. (2022). *Towards a risk-benefit optimized clinical decision support framework for immune-checkpoint inhibitor-treated metastatic melanoma*. University of Wisconsin-Madison.

Huff, D. T., Ferjancic, P., Namías, M., Emamekhoo, H., Perlman, S. B., & Jeraj, R. (2021). Image intensity histograms as imaging biomarkers: Application to immune-related colitis. *Biomedical Physics and Engineering Express*, *7*(6). https://doi.org/10.1088/2057-1976/ac27c3

Huizing, D. M. V., Aalbersberg, E. A., Versleijen, M. W. J., Tesselaar, M. E. T., Walraven, I., Lahaye, M. J., De Wit-Van Der Veen, B. J., & Stokkel, M. P. M. (2020). Early response assessment and prediction of overall survival after peptide receptor radionuclide therapy. *Cancer Imaging*, *20*(1), 1–12. https://doi.org/10.1186/s40644-020-00335-w

Humbert, O., Cadour, N., Paquet, M., Schiappa, R., Poudenx, M., Chardin, D., Borchiellini, D., Benisvy, D., Ouvrier, M. J., Zwarthoed, C., Schiazza, A., Ilie, M., Ghalloussi, H., Koulibaly, P. M., Darcourt, J., & Otto, J. (2020). 18FDG PET/CT in the early assessment of non-small cell lung cancer response to immunotherapy: frequency and clinical significance of atypical evolutive patterns. *European Journal of Nuclear Medicine and Molecular Imaging*, *47*(5), 1158–1167. https://doi.org/10.1007/s00259-019-04573-4

Humbert, O., & Chardin, D. (2020). Dissociated Response in Metastatic Cancer: An Atypical Pattern Brought Into the Spotlight With Immunotherapy. *Frontiers in Oncology*, *10*(September), 1–7. https://doi.org/10.3389/fonc.2020.566297

Huyge, V., Garcia, C., Alexiou, J., Ameye, L., Vanderlinden, B., Lemort, M., Bergmann, P., Awada, A., Body, J. J., & Flamen, P. (2010). Heterogeneity of Metabolic Response to Systemic Therapy in Metastatic Breast Cancer Patients. *Clinical Oncology*, *22*(10), 818–827. https://doi.org/10.1016/j.clon.2010.05.021

Ilan, E., Sandström, M., Wassberg, C., Sundin, A., Garske-Román, U., Eriksson, B., Granberg, D., & Lubberink, M. (2015). Dose response of pancreatic neuroendocrine tumors treated with peptide receptor radionuclide therapy using 177Lu-DOTATATE. *Journal of Nuclear Medicine*, *56*(2), 177–182. https://doi.org/10.2967/jnumed.114.148437

Isensee, F., Jaeger, P. F., Kohl, S. A. A., Petersen, J., & Maier-Hein, K. H. (2021). nnU-Net: a self-configuring method for deep learning-based biomedical image segmentation. *Nature Methods*, *18*(2), 203–211. https://doi.org/10.1038/s41592-020-01008-z

Jaeger, P. F., Kohl, S. A. A., Bickelhaupt, S., Isensee, F., Kuder, T. A., Schlemmer, H.-P., & Maier-Hein, K. H. (2018). *Retina U-Net: Embarrassingly Simple Exploitation of Segmentation Supervision for Medical Object Detection*. http://arxiv.org/abs/1811.08661

Jahn, U., Ilan, E., Sandström, M., Garske-Román, U., Lubberink, M., & Sundin, A. (2020). $^{177}$Lu-DOTATATE Peptide

Receptor Radionuclide Therapy: Dose Response in Small Intestinal Neuroendocrine Tumors. *Neuroendocrinology*, *110*(7–8), 662–670. https://doi.org/10.1159/000504001

Jaqaman, K., Loerke, D., Mettlen, M., Kuwata, H., Grinstein, S., Schmid, S. L., & Danuser, G. (2008). Robust single-particle tracking in live-cell time-lapse sequences. *Nature Methods*, *5*(8), 695–702. https://doi.org/10.1038/nmeth.1237

Jo, I. Y., & Yeo, S. G. (2019). Stereotactic body radiation therapy for pulmonary large cell neuroendocrine carcinoma: A case report. *OncoTargets and Therapy*, *12*, 1359–1364. https://doi.org/10.2147/OTT.S189858

Kamnitsas, K., Ledig, C., Newcombe, V. F. J., Simpson, J. P., Kane, A. D., Menon, D. K., Rueckert, D., & Glocker, B. (2017). Efficient multi-scale 3D CNN with fully connected CRF for accurate brain lesion segmentation. *Medical Image Analysis*, *36*, 61–78. https://doi.org/10.1016/j.media.2016.10.004

Kessler, L. G., Barnhart, H. X., Buckler, A. J., Choudhury, K. R., Kondratovich, M. V., Toledano, A., Guimaraes, A. R., Filice, R., Zhang, Z., & Sullivan, D. C. (2015). The emerging science of quantitative imaging biomarkers terminology and definitions for scientific studies and regulatory submissions. *Statistical Methods in Medical Research*, *24*(1), 9–26. https://doi.org/10.1177/0962280214537333

Kikinis, R., Guttmann, C. R. G., Metcalf, D., Wells, W. M., Ettinger, G. J., Weiner, H. L., & Jolesz, F. A. (1999). Quantitative follow-up of patients with multiple sclerosis using MRI: Technical aspects. *Journal of Magnetic Resonance Imaging*, *9*(4), 519–530. https://doi.org/10.1002/(SICI)1522-2586(199904)9:4<519::AID-JMRI3>3.0.CO;2-M

Klein, S., Staring, M., Murphy, K., Viergever, M. A., & Pluim, J. (2010). elastix: A Toolbox for Intensity-Based Medical Image Registration. *IEEE Transactions on Medical Imaging*, *29*(1), 196–205. https://doi.org/10.1109/TMI.2009.2035616

Klein, S., Staring, M., & Pluim, J. P. W. (2007). Evaluation of optimization methods for nonrigid medical image registration using mutual information and B-splines. *IEEE Transactions on Image Processing*, *16*(12), 2879–2890. https://doi.org/10.1109/TIP.2007.909412

Köhler, C., Wahl, H., Ziemssen, T., Linn, J., & Kitzler, H. H. (2019). Exploring individual multiple sclerosis lesion volume change over time: Development of an algorithm for the analyses of longitudinal quantitative MRI measures. *NeuroImage: Clinical*, *21*(August 2018), 101623. https://doi.org/10.1016/j.nicl.2018.101623

Kratochwil, C., Stefanova, M., Mavriopoulou, E., Holland-Letz, T., Dimitrakopoulou-Strauss, A., Afshar-Oromieh, A., Mier, W., Haberkorn, U., & Giesel, F. L. (2015). SUV of [68Ga]DOTATOC-PET/CT Predicts Response Probability of PRRT in Neuroendocrine Tumors. *Molecular Imaging and Biology*, *17*(3), 313–318. https://doi.org/10.1007/s11307-014-0795-3

Krenning, E., Valkema, R., Kwekkeboom, D. J., Herder, W. W. de, Eijck, C. H. J. van, Jong, M. de, Pauwels, S., & Reubi, J.-C. (2005). Molecular Imaging as In Vivo Molecular Pathology for Gastroenteropancreatic Neuroendocrine Tumors: Implications for Follow-Up After Therapy. *Journal of Nuclear Medicine*, *46*(1).

Kuckertz, S., Klein, J., Engel, C., Geisler, B., Kraß, S., & Heldmann, S. (2022). *Fully automated longitudinal tracking and in-depth analysis of the entire tumor burden: unlocking the complexity*. *April*, 86. https://doi.org/10.1117/12.2613080

Kuckertz, S., Weiler, F., Matusche, B., Lukas, C., Spies, L., Klein, J., & Heldmann, S. (2021). A system for fully automated monitoring of lesion evolution over time in multiple sclerosis. In K. Drukker & M. A. Mazurowski (Eds.), *Medical Imaging 2021: Computer-Aided Diagnosis* (Issue February 2021, p. 83). SPIE. https://doi.org/10.1117/12.2582156

Kulke, M. H., Ajani, J., I, A. B. B. I. I., Byrd, D., Doherty, G. M., Engstrom, P. F., Ettinger, D. S., Gibbs, J. F., Heslin, M. J., Kandeel, F., Kessinger, A., Kulke, M. H., Kvols, L., Jr., A. A. N., Jr., J. A. O., Ratliff, T. W., Saltz, L., Schteingart, D. E., Shah, M. H., & Shibata, S. (2012). Neuroendocrine tumors: Clinical Practice Guidelines in Oncology. *JNCCN Journal of the National Comprehensive Cancer Network*, *4*(2), 102–138.

http://ovidsp.ovid.com/ovidweb.cgi?T=JS&PAGE=reference&D=emed7&NEWS=N&AN=2006403204

Kwekkeboom, D. J., De Herder, W. W., Kam, B. L., Van Eijck, C. H., Van Essen, M., Kooij, P. P., Feelders, R. A., Van Aken, M. O., & Krenning, E. P. (2008). Treatment with the radiolabeled somatostatin analog [177Lu-DOTA0,Tyr3]octreotate: Toxicity, efficacy, and survival. *Journal of Clinical Oncology*, *26*(13), 2124–2130. https://doi.org/10.1200/JCO.2007.15.2553

Kwekkeboom, D. J., Kam, B. L., Van Essen, M., Teunissen, J. J. M., Van Eijck, C. H. J., Valkema, R., De Jong, M., De Herder, W. W., & Krenning, E. P. (2010). Somatostatin receptor-based imaging and therapy of gastroenteropancreatic neuroendocrine tumors. *Endocrine-Related Cancer*, *17*(1). https://doi.org/10.1677/ERC-09-0078

Kyriakopoulos, C. E., Heath, E. I., Ferrari, A., Sperger, J. M., Singh, A., Perlman, S. B., Roth, A. R., Perk, T. G., Modelska, K., Porcari, A., Duggan, W., Lang, J. M., Jeraj, R., & Liu, G. (2020). Exploring Spatial-Temporal Changes in 18F-Sodium Fluoride PET/CT and Circulating Tumor Cells in Metastatic Castration-Resistant Prostate Cancer Treated with Enzalutamide. *Journal of Clinical Oncology*, *38*(31), 3662–3671. https://doi.org/10.1200/JCO.20.00348

Lambin, P., Rios-Velazquez, E., Leijenaar, R., Carvalho, S., Van Stiphout, R. G. P. M., Granton, P., Zegers, C. M. L., Gillies, R., Boellard, R., Dekker, A., & Aerts, H. J. W. L. (2012). Radiomics: Extracting more information from medical images using advanced feature analysis. *European Journal of Cancer*, *48*(4), 441–446. https://doi.org/10.1016/j.ejca.2011.11.036

Larson, S. (1999). Tumor Treatment Response Based on Visual and Quantitative Changes in Global Tumor Glycolysis Using PET-FDG Imaging The Visual Response Score and the Change in Total Lesion Glycolysis. *Clinical Positron Imaging*, *2*(3), 159–171. https://doi.org/10.1016/S1095-0397(99)00016-3

Laudicella, R., Comelli, A., Liberini, V., Vento, A., Stefano, A., Spataro, A., Crocè, L., Baldari, S., Bambaci, M., Deandreis, D., Arico', D., Ippolito, M., Gaeta, M., Alongi, P., Minutoli, F., Burger, I. A., & Baldari, S. (2022). [68Ga]DOTATOC PET/CT Radiomics to Predict the Response in GEP-NETs Undergoing [177Lu]DOTATOC PRRT: The "Theragnomics" Concept. *Cancers*, *14*(4), 984. https://doi.org/10.3390/cancers14040984

Lee, O. N. Y., Tan, K. V., Tripathi, V., Yuan, H., Chan, W. W. L., & Chiu, K. W. H. (2022). The Role of 68Ga-DOTA-SSA PET/CT in the Management and Prediction of Peptide Receptor Radionuclide Therapy Response for Patients With Neuroendocrine Tumors: A Systematic Review and Meta-analysis. *Clinical Nuclear Medicine*, *47*(9), 781–793. https://doi.org/10.1097/RLU.0000000000004235

Lee, S., Wolberg, G., Chwa, K. Y., & Shin, S. Y. (1996). Image metamorphosis with scattered feature constraints. *IEEE Transactions on Visualization and Computer Graphics*, *2*(4), 337–354. https://doi.org/10.1109/2945.556502

Lee, Y., Kim, H. Y., Lee, S. H., Lim, K. Y., Lee, G. K., Yun, T., Han, J. Y., Kim, H. T., & Lee, J. S. (2014). Clinical significance of heterogeneity in response to retreatment with epidermal growth factor receptor tyrosine kinase inhibitors in patients with lung cancer acquiring secondary resistance to the drug. *Clinical Lung Cancer*, *15*(2), 145–151. https://doi.org/10.1016/j.cllc.2013.11.008

Lemieux, L., Wieshmann, U. C., Moran, N. F., Fish, D. R., & Shorvon, S. D. (1998). The detection and significance of subtle changes in mixed-signal brain lesions by serial MRI scan matching and spatial normalization. *Medical Image Analysis*, *2*(3), 227–242. https://doi.org/10.1016/s1361-8415(98)80021-2

Liberini, V., Huellner, M. W., Grimaldi, S., Finessi, M., Thuillier, P., Muni, A., Pellerito, R. E., Papotti, M. G., Piovesan, A., Arvat, E., & Deandreis, D. (2020). The challenge of evaluating response to peptide receptor radionuclide therapy in gastroenteropancreatic neuroendocrine tumors: The present and the future. *Diagnostics*, *10*(12), 1–31. https://doi.org/10.3390/diagnostics10121083

Lin, C., Bradshaw, T., Perk, T., Harmon, S., Eickhoff, J., Jallow, N., Choyke, P. L., Dahut, W. L., Larson, S., Humm, J. L., Perlman, S., Apolo, A. B., Morris, M. J., Liu, G., & Jeraj, R. (2016). Repeatability of quantitative 18F-NaF

PET: A multicenter study. *Journal of Nuclear Medicine*, *57*(12), 1872–1879. https://doi.org/10.2967/jnumed.116.177295

Lou, X., Qin, Y., Xu, X., Yu, X., & Ji, S. (2022). Spatiotemporal heterogeneity and clinical challenge of pancreatic neuroendocrine tumors. *Biochimica et Biophysica Acta - Reviews on Cancer*, *1877*(5), 188782. https://doi.org/10.1016/j.bbcan.2022.188782

Lubner, M., Brace, C., Ziemlewicz, T., Hinshaw, J., & Lee, F. (2013). Microwave Ablation of Hepatic Malignancy. *Seminars in Interventional Radiology*, *30*(01), 056–066. https://doi.org/10.1055/s-0033-1333654

Lyu, N., Kong, Y., Li, X., Mu, L., Deng, H., Chen, H., He, M., Lai, J., Li, J., Tang, H., Lin, Y., & Zhao, M. (2020). Ablation Reboots the Response in Advanced Hepatocellular Carcinoma With Stable or Atypical Response During PD-1 Therapy: A Proof-of-Concept Study. *Frontiers in Oncology*, *10*(October), 1–10. https://doi.org/10.3389/fonc.2020.580241

Maitland, N. J. (2021). Resistance to antiandrogens in prostate cancer: Is it inevitable, intrinsic or induced? *Cancers*, *13*(2), 1–32. https://doi.org/10.3390/cancers13020327

Marstal, K., Berendsen, F., Staring, M., & Klein, S. (2016). SimpleElastix: A User-Friendly, Multi-lingual Library for Medical Image Registration. *2016 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, 574–582. https://doi.org/10.1109/CVPRW.2016.78

Mattes, D., Haynor, D. R., Vesselle, H., Lewellen, T. K., & Eubank, W. (2003). PET-CT image registration in the chest using free-form deformations. *IEEE Transactions on Medical Imaging*, *22*(1), 120–128. https://doi.org/10.1109/TMI.2003.809072

Melloni, G., Gajate, A. M. S., Sestini, S., Gallivanone, F., Bandiera, A., Landoni, C., Muriana, P., Gianolli, L., & Zannini, P. (2013). New positron emission tomography derived parameters as predictive factors for recurrence in resected stage i non-small cell lung cancer. *European Journal of Surgical Oncology*, *39*(11), 1254–1261. https://doi.org/10.1016/j.ejso.2013.07.092

Metcalf, D., Kikinis, R., Guttmann, C., Vaina, L., & Jolesz, F. (1992). 4D connected component labelling applied to quantitative analysis of MS lesion temporal development. *Proceedings of the Annual International Conference of the IEEE Engineering in Medicine and Biology Society, EMBS*, *3*(2), 945–946. https://doi.org/10.1109/IEMBS.1992.5761319

Michielin, O., Atkins, M. B., Koon, H. B., Dummer, R., & Ascierto, P. A. (2020). Evolving impact of long-Term survival results on metastatic melanoma treatment. *Journal for ImmunoTherapy of Cancer*, *8*(2), 1–13. https://doi.org/10.1136/jitc-2020-000948

Mohan, H., Nicholson, P., Winter, D. C., O'Shea, D., O'Toole, D., Geoghegan, J., Maguire, D., Hoti, E., Traynor, O., & Cantwell, C. P. (2015). Radiofrequency ablation for neuroendocrine liver metastases: A systematic review. *Journal of Vascular and Interventional Radiology*, *26*(7), 935-942.e1. https://doi.org/10.1016/j.jvir.2014.12.009

Moltz, J. H., D'Anastasi, M., Kießling, A., Pinto Dos Santos, D., Schülke, C., & Peitgen, H. O. (2012). Workflow-centred evaluation of an automatic lesion tracking software for chemotherapy monitoring by CT. *European Radiology*, *22*(12), 2759–2767. https://doi.org/10.1007/s00330-012-2545-8

Moltz, J. H., Schwier, M., & Peitgen, H. O. (2009). A general framework for automatic detection of matching lesions in follow-up ct. *Proceedings - 2009 IEEE International Symposium on Biomedical Imaging: From Nano to Macro, ISBI 2009*, 843–846. https://doi.org/10.1109/ISBI.2009.5193184

Munkres, J. (1957). Algorithms for the assignment and transportation problems. Journal of the society for industrial and applied mathematics. *J Soc Indust Appl Math*, *5*(1), 32–38.

Namura, K., Minamimoto, R., Yao, M., Makiyama, K., Murakami, T., Sano, F., Hayashi, N., Tateishi, U., Ishigaki, H., Kishida, T., Miura, T., Kobayashi, K., Noguchi, S., Inoue, T., Kubota, Y., & Nakaigawa, N. (2010). Impact of maximum Standardized Uptake Value (SUVmax) evaluated by 18-Fluoro-2-deoxy-D-glucose positron

emission tomography/computed tomography (18F-FDG-PET/CT) on survival for patients with advanced renal cell carcinoma: A preliminary report. *BMC Cancer*, *10*. https://doi.org/10.1186/1471-2407-10-667

O'Connor, J. P. B., Aboagye, E. O., Adams, J. E., Aerts, H. J. W. L., Barrington, S. F., Beer, A. J., Boellaard, R., Bohndiek, S. E., Brady, M., Brown, G., Buckley, D. L., Chenevert, T. L., Clarke, L. P., Collette, S., Cook, G. J., Desouza, N. M., Dickson, J. C., Dive, C., Evelhoch, J. L., … Waterton, J. C. (2017). Imaging biomarker roadmap for cancer studies. *Nature Reviews Clinical Oncology*, *14*(3), 169–186. https://doi.org/10.1038/nrclinonc.2016.162

Ohlendorf, F., Henkenberens, C., Brunkhorst, T., Ross, T. L., Christiansen, H., Bengel, F. M., & Derlin, T. (2020). Volumetric 68Ga-DOTA-TATE PET/CT for assessment of whole-body tumor burden as a quantitative imaging biomarker in patients with metastatic gastroenteropancreatic neuroendocrine tumors. *The Quarterly Journal of Nuclear Medicine and Molecular Imaging*. https://doi.org/10.23736/s1824-4785.20.03238-0

Ohri, N., Tomé, W. A., Méndez Romero, A., Miften, M., Ten Haken, R. K., Dawson, L. A., Grimm, J., Yorke, E., & Jackson, A. (2021). Local Control After Stereotactic Body Radiation Therapy for Liver Tumors. *International Journal of Radiation Oncology Biology Physics*, *110*(1), 188–195. https://doi.org/10.1016/j.ijrobp.2017.12.288

Öksüz, M. Ö., Winter, L., Pfannenberg, C., Reischl, G., Müssig, K., Bares, R., & Dittmann, H. (2014). Peptide receptor radionuclide therapy of neuroendocrine tumors with 90Y-DOTATOC: Is treatment response predictable by pre-therapeutic uptake of 68Ga-DOTATOC? *Diagnostic and Interventional Imaging*, *95*(3), 289–300. https://doi.org/10.1016/j.diii.2013.07.006

Opalińska, M., Morawiec-Sławek, K., Kania-Kuc, A., Al Maraih, I., Sowa-Staszczak, A., & Hubalewska-Dydejczyk, A. (2022). Potential value of pre- and post-therapy [68Ga]Ga-DOTA-TATE PET/CT in the prognosis of response to PRRT in disseminated neuroendocrine tumors. *Frontiers in Endocrinology*, *13*(August), 1–10. https://doi.org/10.3389/fendo.2022.929391

Ortega, C., Wong, R. K. S., Schaefferkoetter, J., Veit-Haibach, P., Myrehaug, S., Juergens, R., Laidley, D., Anconina, R., Liu, A., & Metser, U. (2021). Quantitative 68Ga-Dotatate PET/CT parameters for the prediction of therapy response in patients with progressive metastatic neuroendocrine tumors treated with 177Lu-Dotatate. *Journal of Nuclear Medicine*, *62*(10), jnumed.120.256727. https://doi.org/10.2967/jnumed.120.256727

Otsu, N. (1979). A thresholding selection method from graylevel histogram. *IEEE Trans. Syst., Man, Cybern.*, *9*(1), 62–66.

Owonikoko, T. K., Kim, H. R., Govindan, R., Ready, N., Reck, M., Peters, S., Dakhil, S. R., Navarro, A., Rodriguez-Cid, J., Schenker, M., Lee, J. S., Gutierrez, V., Percent, I., Morgensztern, D., Fairchild, J., Baudelet, C., & Park, K. (2019). Nivolumab (nivo) plus ipilimumab (ipi), nivo, or placebo (pbo) as maintenance therapy in patients (pts) with extensive disease small cell lung cancer (ED-SCLC) after first-line (1L) platinum-based chemotherapy (chemo): Results from the double-blind, rando. *Annals of Oncology*, *30*(Supplement 2), ii77. https://doi.org/10.1093/annonc/mdz094

Palma, D. A., Salama, J. K., Lo, S. S., Senan, S., Treasure, T., Govindan, R., & Weichselbaum, R. (2014). The oligometastatic state-separating truth from wishful thinking. *Nature Reviews Clinical Oncology*, *11*(9), 549–557. https://doi.org/10.1038/nrclinonc.2014.96

Pan, L. L., Gu, P., Huang, G., Xue, H. P., & Wu, S. Q. (2009). Prognostic significance of SUV on PET/CT in patients with esophageal cancer: A systematic review and meta-analysis. *European Journal of Gastroenterology and Hepatology*, *21*(9), 1008–1015. https://doi.org/10.1097/MEG.0b013e328323d6fa

Park, M. S., & Lee, S. M. (2014). Preoperative 18 F-FDG PET-CT maximum standardized uptake value predicts recurrence of biliary tract cancer. *Anticancer Research*, *34*(5), 2551–2554.

Pauwels, E., Van Binnebeek, S., Vandecaveye, V., Baete, K., Vanbilloen, H., Koole, M., Mottaghy, F. M., Haustermans, K., Clement, P. M., Nackaerts, K., Van Cutsem, E., Verslype, C., & Deroose, C. M. (2020). Inflammation-based index and 68ga-dotatoc pet-derived uptake and volumetric parameters predict outcome in neuroendocrine tumor patients treated with 90y-dotatoc. *Journal of Nuclear Medicine*, *61*(7), 1014–1020. https://doi.org/10.2967/jnumed.119.236935

Pauwels, S., Barone, R., Walrand, S., Borson-Chazot, F., Valkema, R., Kvols, L. K., Krenning, E. P., & Jamar, F. (2005). Practical dosimetry of peptide receptor radionuclide therapy with 90Y-labeled somatostatin analogs. *Journal of Nuclear Medicine*, *46*(1 SUPPL.), 92–98.

Peng, H., Long, F., & Ding, C. (2005). Feature selection based on mutual information: Criteria of Max-Dependency, Max-Relevance, and Min-Redundancy. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *27*(8), 1226–1238. https://doi.org/10.1109/TPAMI.2005.159

Perk, T., Chen, S., Harmon, S., Lin, C., Bradshaw, T., Perlman, S., Liu, G., & Jeraj, R. (2018). A statistically optimized regional thresholding method (SORT) for bone lesion detection in 18F-NaF PET/CT imaging. *Physics in Medicine and Biology*, *63*(22), 225018. https://doi.org/10.1088/1361-6560/aaebba

Perk, T., Yip, S., Jeraj, R., & Perlman, S. (2020). Spatiotemporal evolution of lesion response heterogeneity to 177Lu-DOTATATE therapy in advanced gastroenteropancreatic neuroendocrine tumors. *Journal of Nuclear Medicine*, *61*(supplement 1), 1132. https://jnm.snmjournals.org/content/61/supplement_1/1132

Pirasteh, A., Lovrec, P., & Bodei, L. (2021). Imaging of neuroendocrine tumors: A pictorial review of the clinical value of different imaging modalities. *Reviews in Endocrine and Metabolic Disorders*, *22*(3), 539–552. https://doi.org/10.1007/s11154-021-09631-7

Ralli, M., Botticelli, A., Visconti, I. C., Angeletti, D., Fiore, M., Marchetti, P., Lambiase, A., De Vincentiis, M., & Greco, A. (2020). Immunotherapy in the Treatment of Metastatic Melanoma: Current Knowledge and Future Directions. *Journal of Immunology Research*, *2020*. https://doi.org/10.1155/2020/9235638

Roll, W., Weckesser, M., Seifert, R., Bodei, L., & Rahbar, K. (2021). Imaging and liquid biopsy in the prediction and evaluation of response to PRRT in neuroendocrine tumors: implications for patient management. *European Journal of Nuclear Medicine and Molecular Imaging*, 4016–4027. https://doi.org/10.1007/s00259-021-05359-3

Roth, A., Cooley, G., Smilowitz, J., Ferjancic, P., Liu, G., & Jeraj, R. (2020). Treatment of Oligoresistant and Oligoprogressive Disease in Metastatic Prostate Cancer Patients with Radiation Therapy. *AAPM Meeting*.

Rueckert, D., Sonoda, L. I., Hayes, C., Hill, D. L. G., Leach, M. O., & Hawkes, D. J. (1999). Nonrigid registration using free-form deformations: application to breast MR images. *IEEE Transactions on Medical Imaging*, *18*(8), 18:712-21.

Russo, M., Siravegna, G., Blaszkowsky, L. S., Corti, G., Crisafulli, G., Ahronian, L. G., Mussolin, B., Kwak, E. L., Buscarino, M., Lazzari, L., Valtorta, E., Truini, M., Jessop, N. A., Robinson, H. E., Hong, T. S., Mino-Kenudson, M., Di Nicolantonio, F., Thabet, A., Sartore-Bianchi, A., … Corcoran, R. B. (2016). Tumor heterogeneity and Lesion-Specific response to targeted therapy in colorectal cancer. *Cancer Discovery*, *6*(2), 147–153. https://doi.org/10.1158/2159-8290.CD-15-1283

S., L. L., & Wahba, G. (2006). Spline Models for Observational Data. *Mathematics of Computation*, *57*(195), 444. https://doi.org/10.2307/2938687

Saenger, Y. M., & Wolchok, J. D. (2008). The heterogeneity of the kinetics of response to ipilimumab in metastatic melanoma: Patient cases. *Cancer Immunity*, *8*(January), 1–7.

Sahani, D. V., Bonaffini, P. A., Fernández-Del Castillo, C., & Blake, M. A. (2013). Gastroenteropancreatic neuroendocrine tumors: Role of imaging in diagnosis and management. *Radiology*, *266*(1), 38–61. https://doi.org/10.1148/radiol.12112512

Sainz-Esteban, A., Prasad, V., Schuchardt, C., Zachert, C., Carril, J. M., & Baum, R. P. (2012). Comparison of

sequential planar 177Lu-DOTA-TATE dosimetry scans with 68Ga-DOTA-TATE PET/CT images in patients with metastasized neuroendocrine tumours undergoing peptide receptor radionuclide therapy. *European Journal of Nuclear Medicine and Molecular Imaging*, *39*(3), 501–511. https://doi.org/10.1007/s00259-011-2003-x

Santoro-Fernandes, V., Huff, D., Scarpelli, M. L., Perk, T. G., Albertini, M. R., Perlman, S., Yip, S. S. F. F., & Jeraj, R. (2021). Development and validation of a longitudinal soft-tissue metastatic lesion matching algorithm. *Physics in Medicine & Biology*, *66*(15), 155017. https://doi.org/10.1088/1361-6560/ac1457

Scarpelli, M., Bruce, J. Y., Carmichael, L., Eickhoff, J., Kolesar, J., Perlman, S., Jeraj, R., & Liu, G. (2016). 18F-FLT PET/CT imaging in patients with advanced solid malignancies treated with axitinib on an intermittent dosing regimen. *Cancer Chemotherapy and Pharmacology*, *78*(6), 1245–1252. https://doi.org/10.1007/s00280-016-3183-7

Scarpelli, M., Simoncic, U., Perlman, S., Liu, G., & Jeraj, R. (2018). Dynamic 18 F-FLT PET imaging of spatiotemporal changes in tumor cell proliferation and vasculature reveals the mechanistic actions of anti-angiogenic therapy. *Physics in Medicine & Biology*, *63*(15), 155008. https://doi.org/10.1088/1361-6560/aad1be

Schott, B., Weisman, A. J., Perk, T. G., Roth, A. R., Liu, G., & Jeraj, R. (2023). Comparison of automated full-body bone metastases delineation methods and their corresponding prognostic power. *Physics in Medicine and Biology*, *68*(3). https://doi.org/10.1088/1361-6560/acaf22

Schwartz, L. H., Mazumdar, M., Brown, W., Smith, A., & Panicek, D. M. (2003). Variability in response assessment in solid tumors: Effect of number of lesions chosen for measurement. *Clinical Cancer Research*, *9*(12), 4318–4323.

Schwarz, J. D., Bader, M., Jenicke, L., Hemminger, G., Jänicke, F., & Avril, N. (2005). Early prediction of response to chemotherapy in metastatic breast cancer using sequential18F-FDG PET. *Journal of Nuclear Medicine*, *46*(7), 1144–1150.

Sederberg, T. W., & Parry, S. R. (1986). Free-form deformation of solid geometric models. *Proceedings of the 13th Annual Conference on Computer Graphics and Interactive Techniques, SIGGRAPH 1986*, *20*(4), 151–160. https://doi.org/10.1145/15922.15903

Seifert, R., Herrmann, K., Kleesiek, J., Schäfers, M., Shah, V., Xu, Z., Chabin, G., Grbic, S., Spottiswoode, B., & Rahbar, K. (2020). Semi-automatically quantified tumor volume using Ga-68-PSMA-11-PET as biomarker for survival in patients with advanced prostate cancer. *Journal of Nuclear Medicine*, 1–30. https://doi.org/10.2967/jnumed.120.242057

Seifert, R., Kessel, K., Schlack, K., Weber, M., Herrmann, K., Spanke, M., Fendler, W. P., Hadaschik, B., Kleesiek, J., Schäfers, M., Weckesser, M., Boegemann, M., & Rahbar, K. (2021). PSMA PET total tumor volume predicts outcome of patients with advanced prostate cancer receiving [177Lu]Lu-PSMA-617 radioligand therapy in a bicentric analysis. *European Journal of Nuclear Medicine and Molecular Imaging*, *48*(4), 1200–1210. https://doi.org/10.1007/s00259-020-05040-1

Sgouros, G., Dewaraja, Y. K., Escorcia, F., Graves, S. A., Hope, T. A., Iravani, A., Pandit-Taskar, N., Saboury, B., James, S. S., & Zanzonico, P. B. (2021). Tumor Response to Radiopharmaceutical Therapies: The Knowns and the Unknowns. *Journal of Nuclear Medicine : Official Publication, Society of Nuclear Medicine*, *62*(12), 12S-22S. https://doi.org/10.2967/jnumed.121.262750

Shafiei, A., Bagheri, M., Farhadi, F., Apolo, A. B., Biassou, N. M., Folio, L. R., Jones, E. C., & Summers, R. M. (2021). CT Evaluation of Lymph Nodes That Merge or Split during the Course of a Clinical Trial: Limitations of RECIST 1.1. *Radiology: Imaging Cancer*, *3*(3), e200090. https://doi.org/10.1148/rycan.2021200090

Shahar, A., & Greenspan, H. (2005). A probabilistic framework for the detection and tracking in time of multiple sclerosis lesions. *2004 2nd IEEE International Symposium on Biomedical Imaging: Nano to Macro (IEEE Cat*

*No. 04EX821)*, 440–443. https://doi.org/10.1109/isbi.2004.1398569

Shamonin, D. P., Bron, E. E., Lelieveldt, B. P. F., Smits, M., Klein, S., & Staring, M. (2014). Fast parallel image registration on CPU and GPU for diagnostic classification of Alzheimer's disease. *Frontiers in Neuroinformatics*, *7*(JAN), 1–15. https://doi.org/10.3389/fninf.2013.00050

Sharabi, A., Kim, S. S., Kato, S., Sanders, P. D., Patel, S. P., Sanghvi, P., Weihe, E., & Kurzrock, R. (2017). Exceptional Response to Nivolumab and Stereotactic Body Radiation Therapy (SBRT) in Neuroendocrine Cervical Carcinoma with High Tumor Mutational Burden: Management Considerations from the Center For Personalized Cancer Therapy at UC San Diego Moores Cance. *The Oncologist*, *22*(6), 631–637. https://doi.org/10.1634/theoncologist.2016-0517

Sharma, R., Wang, W. M., Yusuf, S., Evans, J., Ramaswami, R., Wernig, F., Frilling, A., Mauri, F., Al-Nahhas, A., Aboagye, E. O., & Barwick, T. D. (2019). 68Ga-DOTATATE PET/CT parameters predict response to peptide receptor radionuclide therapy in neuroendocrine tumours. *Radiotherapy and Oncology*, *141*, 108–115. https://doi.org/10.1016/j.radonc.2019.09.003

Sotiras, A., Davatzikos, C., & Paragios, N. (2014). Deformable Medical Image Registration: A Survey. *IEEE Trans Med Imaging*, *32*(7), 1153–1190. https://doi.org/10.1109/TMI.2013.2265603.Deformable

Soydal, Ç., Peker, A., Özkan, E., Küçük, Ö. N., & Kir, M. K. (2016). The role of baseline Ga-68 DOTATATE positron emission tomography/computed tomography in the prediction of response to fixed-dose peptide receptor radionuclide therapy with lu-177 DOTATATE. *Turkish Journal of Medical Sciences*, *46*(2), 409–413. https://doi.org/10.3906/sag-1412-11

Stehman, V. S. (1997). Selecting and interpreting measures of thematic classification accuracy. *Remote Sensing of Environment*, *62*, 77.

Strosberg, J., Al-Toubah, T., & Cives, M. (2023). In Defense of Neuroendocrine Tumor Trials. *JAMA Oncology*. https://doi.org/10.1001/jamaoncol.2023.3392

Strosberg, J., El-Haddad, G., Wolin, E., Hendifar, A., Yao, J., Chasen, B., Mittra, E., Kunz, P. L., Kulke, M. H., Jacene, H., Bushnell, D., O'Dorisio, T. M., Baum, R. P., Kulkarni, H. R., Caplin, M., Lebtahi, R., Hobday, T., Delpassand, E., Van Cutsem, E., … Krenning, E. (2017). Phase 3 Trial of 177 Lu-Dotatate for Midgut Neuroendocrine Tumors . *New England Journal of Medicine*, *376*(2), 125–135. https://doi.org/10.1056/nejmoa1607427

Studholme, C., Hawkes, D. J., & Hill, D. L. (1998). Normalized entropy measure for multimodality image alignment. In K. M. Hanson (Ed.), *Medical Imaging 1998: Image Processing* (Vol. 3338, Issue June 1998, pp. 132–143). https://doi.org/10.1117/12.310835

Sun, W., & Schuchter, L. M. (2001). Metastatic melanoma. *Current Treatment Options in Oncology*, *1*, 193–202.

Sundin, A., & Rockall, A. (2012). Therapeutic monitoring of gastroenteropancreatic neuroendocrine tumors: The challenges ahead. *Neuroendocrinology*, *96*(4), 261–271. https://doi.org/10.1159/000342270

Switzer, B., Puzanov, I., Skitzki, J. J., Hamad, L., & Ernstoff, M. S. (2022). Managing Metastatic Melanoma in 2022: A Clinical Review. *JCO Oncology Practice*, *18*(5), 335–351. https://doi.org/10.1200/op.21.00686

Szeskin, A., Rochman, S., Weiss, S., Lederman, R., Sosna, J., & Joskowicz, L. (2023). Liver lesion changes analysis in longitudinal CECT scans by simultaneous deep learning voxel classification with SimU-Net. *Medical Image Analysis*, *83*(February 2022), 102675. https://doi.org/10.1016/j.media.2022.102675

Tazdait, M., Mezquita, L., Lahmar, J., Ferrara, R., Bidault, F., Ammari, S., Balleyguier, C., Planchard, D., Gazzah, A., Soria, J. C., Marabelle, A., Besse, B., & Caramella, C. (2018). Patterns of responses in metastatic NSCLC during PD-1 or PDL-1 inhibitor therapy: Comparison of RECIST 1.1, irRECIST and iRECIST criteria. *European Journal of Cancer*, *88*, 38–47. https://doi.org/10.1016/j.ejca.2017.10.017

Teker, F., & Elboga, U. (2021). Is SUVmax a useful marker for progression-free survival in patients with

metastatic GEP-NET receiving 177Lu-DOTATATE therapy? *Hellenic Journal of Nuclear Medicine*, *24*(2), 122–131. https://doi.org/10.1967/s002449912352

Therasse, P., Arbuck, S. G., Eisenhauer, E. A., Wanders, J., Kaplan, R. S., Rubinstein, L., Verweij, J., Van Glabbeke, M., Van Oosterom, A. T., Christian, M. C., & Gwyther, S. G. (2000). New guidelines to evaluate the response to treatment in solid tumors. *Journal of the National Cancer Institute*, *92*(3), 205–216. https://doi.org/10.1093/jnci/92.3.205

Tozuka, T., Kitazono, S., Sakamoto, H., Yoshida, H., Amino, Y., Uematsu, S., Yoshizawa, T., Hasegawa, T., Uchibori, K., Yanagitani, N., Horiike, A., Horai, T., Seike, M., Gemma, A., & Nishio, M. (2020). Dissociated responses at initial computed tomography evaluation is a good prognostic factor in non-small cell lung cancer patients treated with anti-programmed cell death-1/ligand 1 inhibitors. *BMC Cancer*, *20*(1), 1–8. https://doi.org/10.1186/s12885-020-6704-z

Urso, L., Castello, A., Treglia, G., Panareo, S., Nieri, A., Rambaldi, I., Caracciolo, M., Ortolan, N., Uccelli, L., Cittanti, C., Castellani, M., & Bartolomei, M. (2023). Relevance of Volumetric Parameters Applied to [68Ga]Ga-DOTATOC PET/CT in NET Patients Treated with PRRT. *Diagnostics*, *13*(4), 1–11. https://doi.org/10.3390/diagnostics13040606

Van Helden, E. J., Hoekstra, O. S., Boellaard, R., Roth, C., Mulder, E. R., Verheul, H. M. W., & Menke-Van Der Houven Van Oordt, C. W. (2016). Early 18F-FDG PET/CT evaluation shows heterogeneous metabolic responses to anti-EGFR therapy in patients with metastatic colorectal cancer. *PLoS ONE*, *11*(5). https://doi.org/10.1371/journal.pone.0155178

Vanderhoek, M., Perlman, S. B., & Jeraj, R. (2013). Impact of Different Standardized Uptake Value Measures on PET-Based Quantification of Treatment Response. *Journal of Nuclear Medicine*, *54*(8), 1188–1194. https://doi.org/10.2967/jnumed.112.113332

Vittinghoff, E., & McCulloch, C. E. (2007). Relaxing the rule of ten events per variable in logistic and cox regression. *American Journal of Epidemiology*, *165*(6), 710–718. https://doi.org/10.1093/aje/kwk052

Wahl, R. L., Jacene, H., Kasamon, Y., & Lodge, M. A. (2009). From RECIST to PERCIST: Evolving Considerations for PET Response Criteria in Solid Tumors. *Journal of Nuclear Medicine*, *50*(Suppl_1), 122S-150S. https://doi.org/10.2967/jnumed.108.057307

Walentowicz-Sadlecka, M., Malkowski, B., Walentowicz, P., Sadlecki, P., Marszalek, A., Pietrzak, T., & Grabiec, M. (2014). The preoperative maximum standardized uptake value measured by 18F-FDG PET/CT as an independent prognostic factor of overall survival in endometrial cancer patients. *BioMed Research International*, *2014*. https://doi.org/10.1155/2014/234813

Weber, M., Kessler, L., Schaarschmidt, B., Fendler, W. P., Lahner, H., Antoch, G., Umutlu, L., Herrmann, K., & Rischpler, C. (2020). Treatment-related changes in neuroendocrine tumors as assessed by textural features derived from 68Ga-DOTATOC PET/MRI with simultaneous acquisition of apparent diffusion coefficient. *BMC Cancer*, *20*(1), 1–12. https://doi.org/10.1186/s12885-020-06836-y

Wehrend, J., Silosky, M., Xing, F., & Chin, B. B. (2021). Automated liver lesion detection in 68Ga DOTATATE PET/CT using a deep fully convolutional neural network. *EJNMMI Research*, *11*(1). https://doi.org/10.1186/s13550-021-00839-x

Weisman, A. J., Kieler, M. W., Perlman, S. B., Hutchings, M., Jeraj, R., Kostakoglu, L., & Bradshaw, T. J. (2020a). Convolutional neural networks for automated pet/ct detection of diseased lymph node burden in patients with lymphoma. *Radiology: Artificial Intelligence*, *2*(5), 1–2. https://doi.org/10.1148/ryai.2020200016

Weisman, A. J., Kieler, M. W., Perlman, S., Hutchings, M., Jeraj, R., Kostakoglu, L., & Bradshaw, T. J. (2020b). Comparison of 11 automated PET segmentation methods in lymphoma. *Physics in Medicine and Biology*, *65*(23). https://doi.org/10.1088/1361-6560/abb6bd

Weisman, A., Lokre, O., Schott, B., Santoro-Fernandes, V., Jeraj, R., Perk, T., Cho, S., & Perlman, S. (2022).

Automated detection and quantification of neuroendocrine tumors on 68Ga-DOTATATE PET/CT images using a U-net ensemble method. *Society of Nuclear Medicine and Molecular Imaging Annual Meeting*.

Weissleder, R. (2006). Molecular imaging in cancer. *Science*, *312*(5777), 1168–1171. https://doi.org/10.1126/science.1125949

Werner, R. A., Ilhan, H., Lehner, S., Papp, L., Zsótér, N., Schatka, I., Muegge, D. O., Javadi, M. S., Higuchi, T., Buck, A. K., Bartenstein, P., Bengel, F., Essler, M., Lapa, C., & Bundschuh, R. A. (2019). Pre-therapy Somatostatin Receptor-Based Heterogeneity Predicts Overall Survival in Pancreatic Neuroendocrine Tumor Patients Undergoing Peptide Receptor Radionuclide Therapy. *Molecular Imaging and Biology*, *21*(3), 582–590. https://doi.org/10.1007/s11307-018-1252-5

Werner, R. A., Lapa, C., Ilhan, H., Higuchi, T., Buck, A. K., Lehner, S., Bartenstein, P., Bengel, F., Schatka, I., Muegge, D. O., Papp, L., Zsótér, N., Große-Ophoff, T., Essler, M., & Bundschuh, R. A. (2017). Survival prediction in patients undergoing radionuclide therapy based on intratumoral somatostatin-receptor heterogeneity. *Oncotarget*, *8*(4), 7039–7049. https://doi.org/10.18632/oncotarget.12402

Xu, J., Greenspan, H., Napel, S., & Rubin, D. L. (2011). Automated temporal tracking and segmentation of lymphoma on serial CT examinations. *Medical Physics*, *38*(11), 5879–5886. https://doi.org/10.1118/1.3643027

Yan, K., Wang, X., Lu, L., Zhang, L., Harrison, A. P., Bagheri, M., & Summers, R. M. (2018). Deep Lesion Graphs in the Wild: Relationship Learning and Organization of Significant Radiology Image Findings in a Diverse Large-Scale Lesion Database. *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 9261–9270. https://doi.org/10.1109/CVPR.2018.00965

Yang, D., Zheng, J., Nofal, A., Deasy, J., & Naqa, I. El. (2009). Techniques and software tool for 3D multimodality medical image segmentation. *Journal of Radiation Oncology Informatics*, *1*(1), 1–22. https://doi.org/https://doi.org/10.5166/jroi-1-1-4

Yin, Y., Hoffman, E. A., Ding, K., Reinhardt, J. M., & Lin, C. L. (2011). A cubic B-spline-based hybrid registration of lung CT images for a dynamic airway geometric model with large deformation. *Physics in Medicine and Biology*, *56*(1), 203–218. https://doi.org/10.1088/0031-9155/56/1/013

Yip, S., & Jeraj, R. (2014). Use of articulated registration for response assessment of individual metastatic bone lesions. *Physics in Medicine and Biology*, *59*(6), 1501–1514. https://doi.org/10.1088/0031-9155/59/6/1501

Yip, S., Perk, T., & Jeraj, R. (2014). Development and evaluation of an articulated registration algorithm for human skeleton registration. *Physics in Medicine and Biology*, *59*(6), 1485–1499. https://doi.org/10.1088/0031-9155/59/6/1485

Yu, E. Y., Duan, F., Muzi, M., Deng, X., Chin, B. B., Alumkal, J. J., Taplin, M. E., Taub, J. M., Herman, B., Higano, C. S., Doot, R. K., Hartfeil, D., Febbo, P. G., & Mankoff, D. A. (2015). Castration-resistant prostate cancer bone metastasis response measured by 18F-fluoride PET after treatment with dasatinib and correlation with progression-free survival: Results from American College of Radiology Imaging Network 6687. *Journal of Nuclear Medicine*, *56*(3), 354–360. https://doi.org/10.2967/jnumed.114.146936

Zwirtz, K., Hardt, J., Acker, G., Baur, A. D. J., Pavel, M., Huang, K., Brenner, W., & Prasad, V. (2022). Comparison of Choi, RECIST and Somatostatin Receptor PET/CT Based Criteria for the Evaluation of Response and Response Prediction to PRRT. *Pharmaceutics*, *14*(6). https://doi.org/10.3390/pharmaceutics14061278