Evaluating Multi-Modal Data Fusion Approaches for Predictive Clinical Models Using Multiple Medical Data Domains

Ehsan Alipour

A dissertation

submitted in partial fulfillment of the

requirements for the degree of

Doctor of Philosophy

University of Washington

2025

Reading Committee:

Peter Tarczy-Hornoch

Jennifer Hadlock

Majid Chalian

Program Authorized to Offer Degree:

Department of Biomedical Informatics and Medical Education

University of Washington

Abstract

Evaluating Multi-Modal Data Fusion Approaches for Predictive Clinical Models Using Multiple Medical Data Domains

Ehsan Alipour

Chair of the Supervisory Committee:

Peter Tarczy-Hornoch

Department of Biomedical Informatics and Medical Education

Disease outcome prediction is a central research focus in biomedical informatics, as it facilitates precision health related interventions and scientific discovery by enabling digital clinical trials and multiple other benefits. Multimodal deep learning models have emerged as powerful tools in biomedical research, offering the ability to integrate diverse data sources such as clinical records, multi-omics data, imaging, survey responses, and wearable data to enhance predictive accuracy and deepen understanding of medical phenomena. Central to multimodal modeling is the process of *data fusion*, where information from different modalities is integrated into a unified model. Three primary fusion strategies exist in deep learning: early fusion (feature-level), intermediate fusion and late fusion (decision-level). While widely adopted in other domains, their comparative

performance and implementation considerations remain underexplored in biomedical applications, where data heterogeneity, missingness, and varying dimensionality present additional challenges.

This dissertation aims to evaluate the implications of data fusion strategies for developing multimodal predictive models in medicine. Across three distinct aims, I assess the impact of early, intermediate, and late fusion techniques on predictive performance, implementation complexity, and generalizability using diverse combinations of data types, outcomes, and modeling strategies. These studies span multiple datasets and outcome types (binary categorial variables vs continuous ratio variables) providing a broad view of fusion strategy utility in real-world biomedical settings.

In **Aim 1—***Evaluation and comparison of early, intermediate, and late fusion techniques for combining exposures, clinical and genomics data for disease risk prediction task using All of Us: Risk of CKD in patients with type 2 diabetes*—I evaluated and compared early, intermediate, and late fusion strategies for integrating longitudinal EHR, genomic, and survey data to predict chronic kidney disease (CKD) progression in patients with type 2 diabetes using a novel transformer-based multimodal architecture. Using data from the NIH's All of Us initiative, I trained models on a cohort of approximately 40,000 patients. While the best performing unimodal model achieved a baseline performance with an AUROC of 0.73 (0.71 - 0.75), the inclusion of multimodal data offered only marginal improvement with an AUROC of 0.74 (0.72 – 0.76), with the benefit limited to the early fusion approach and lacking statistical significance. This aim highlighted the challenges of integrating multimodal data with different dimensions using transformer models and emphasized the role of modality-specific relative predictive strength.

In **Aim 2—***Development and assessment of the incremental value of combining a deep convolutional neural network feature extractor on imaging data and clinical data on a binary prediction task: Predict post-surgical margin status in soft tissue sarcoma*—I extended the fusion

analysis to imaging data by combining a convolutional neural network (CNN) trained on longitudinal cross-sectional imaging with a shallow neural network trained on clinical and pathology variables to predict post-surgical margin status in patients with soft tissue sarcoma (n=202). Here, the intermediate fusion strategy significantly outperformed other approaches, achieving an AUROC of 0.80 (0.66–0.95), suggesting that cross-modal interactions between histologic features and imaging embeddings may be best captured through intermediate fusion. This result demonstrated the potential value of intermediate fusion when complementary signals exist across modalities.

In **Aim 3—***Evaluation and comparison of early, intermediate, and late fusion techniques for combining imaging and clinical data on a regression prediction task: Estimation of CT-based body composition metrics from chest radiographs***—**I explored fusion strategies for estimating continuous CT-derived body composition metrics (e.g., visceral, and subcutaneous fat volumes) using only chest radiographs and clinical variables in a dataset of 1,088 patients. A multitask multimodal model was developed and evaluated across early, intermediate, and late fusion strategies. Late fusion consistently delivered the best performance across most body composition metrics, closely followed by intermediate fusion. These results suggest that when individual modalities offer high independent predictive power, decision-level integration may be optimal for regression tasks.

Collectively, these aims provide a broad evaluation of data fusion strategies in multimodal biomedical modeling, highlighting their strengths, limitations, and practical considerations. Findings suggest that no single fusion strategy universally outperforms the others; rather, optimal fusion depends on data characteristics, model architecture, and task-specific objectives. This

dissertation lays the groundwork for future research aimed at developing adaptive fusion strategies tailored to the complexities of real-world biomedical data.

Acknowledgments

I would like to express my deepest gratitude to my incredible partner, Atefe Pooyan, whose unwavering support and inspiration have carried me through every stage of this journey. Her encouragement pushed me to aim higher, and her presence sustained me during the most challenging times.

I am deeply thankful to my parents, Mohammad Alipour and Parinoush Kalantari, who instilled in me resilience and a deep appreciation for the value of science.

I am profoundly grateful to my PhD advisor and committee chair, Dr. Peter Tarczy-Hornoch, for his steadfast support and invaluable guidance. I truly cherished every moment of working with him. I would also like to thank Dr. Jennifer Hadlock, co-chair of my committee, whose mentorship was instrumental in shaping the direction of my research and who supported me throughout every step of the process. I am sincerely thankful to Dr. Majid Chalian, my research assistantship advisor and committee member, for believing in me from the start. His support and the opportunities he provided in the musculoskeletal radiology lab were essential in the development of my research. I would also like to thank Dr. Sara Mostafavi and Dr. Marco Carone, members of my committee, for their continued encouragement and support.

Special thanks to the members of the University of Washington Musculoskeletal Radiology Lab, including Dr. Chankue Park, Dr. Sara Haseli, and Dr. Firoozeh Shomal Zadeh, for their contributions to curating the soft tissue sarcoma dataset used in Aim 2. I am also grateful to members of the Hadlock Lab at the Institute for Systems Biology—Bhargav Vemuri, Qi Wei, and Alexandra Ralevski—for their collaboration. I would like to extend my thanks to the Truveta

Research team, especially Dr. Nickolas Stucky and Sam Gratzl, for their partnership on Aim 3 of my dissertation. Additionally, I am thankful to the members of the Precision Medicine Informatics Group (PMIG), from whom I learned a great deal.

Finally, I am honored to have been part of the BIME family. I extend my sincere thanks to Dr. John Gennari, director of our graduate program, for his leadership and support.

Table of Contents

1. Executive Summary

1.1 Overview

The use of artificial intelligence to develop predictive medical models has introduced many opportunities in medicine. Disease outcome prediction has become a central research focus in biomedical informatics, as it can facilitate precision health related interventions and scientific discovery by enabling digital clinical trials and multiple other benefits. Over the recent years, multimodal deep learning has become increasingly popular for developing predictive medical models. These models have the advantage of utilizing data from difference sources and extracting intra- and cross-modal relationships that can result in a better understanding of the patient and their disease trajectory. Despite its benefits, multimodal modeling introduces challenges including variability in data size and dimensionality, variable missingness levels, and diverse data type specific analytic methods. The process of integrating multimodal data into a deep learning model is called *data fusion.* The three main approaches to data fusion in deep learning include *early fusion (feature-level fusion), intermediate fusion,* and *late fusion (decision-level fusion).* Early fusion includes concatenation of raw variables or features extracted from raw variables before passing the data through the deep learning network. Intermediate fusion involves the use of modality specific networks to extract embeddings from each modality that are then combined and passed through the rest of the network. In late fusion, separate unimodal models are trained for each modality. Subsequently, the predictions of these models are combined using voting strategies or a small separate model to make the final prediction.

The implications of different fusion strategies have not been extensively studied in biomedical applications. Few studies have explored different fusion approaches and compared their performance. In addition, these studies have found conflicting results regarding the best fusion

strategy. Currently, it is recommended that researchers treat fusion strategy as a hyperparameter. Given this gap in the literature, in this body of work, my overarching question was to evaluate the implications of multimodal data fusion approaches in developing predictive medical models using examples from multiple medical data domains. I explored fusion of longitudinal EHR, clinical, pathology, imaging, genomics, and survey responses data using various deep learning network architectures including transformers, convolutional neural networks, fully connected neural networks and tree-based methods. I explored the implications of fusion strategies both with respect to performance metrics and implementation considerations and challenges attributed each fusion strategy.

1.2 Motivation

Given the potential for multimodal predictive medical models to improve individualized care and drive scientific discovery, the work of this dissertation was motivated by the goal to evaluate and compare early, intermediate, and late fusion strategies for developing deep learning predictive models using multiple medical data domains, both in terms of final performance metrics and implications for multimodal modeling with respect to implementation challenges, resources required, and data types or dataset characteristics that favor a specific type of data fusion. Throughout the dissertation, I explored these fusion strategies in three settings: 1) combining longitudinal EHR, genomics and survey data using transformer models for binary outcome prediction, 2) combining longitudinal, cross-sectional imaging data with clinical and pathology variables, and 3) combining 2D imaging and clinical variables for regression tasks. As a first step, I explored prediction of progression to CKD in patients with type 2 diabetes using a multimodal transformer-based architecture to combine EHR, genomic and survey responses data using early, intermediate, and late fusion.

1.3 Research Aims

My overarching question is *evaluation of multimodal data fusion approaches in predictive models using medical data from multiple domains*. My dissertation includes three aims, each of which explores fusion in the setting of a different biomedical question, using different biomedical data types and outcomes. These include: 1) Evaluation and comparison of early, intermediate, and late fusion techniques for combining exposures, clinical and genomics data for disease risk prediction task using NIH All of Us (1): Risk of CKD in patients with type 2 diabetes; 2) Development and assessment of the incremental value of combining a deep convolutional neural network feature extractor on imaging data and clinical data on a binary prediction task: Predict post-surgical margin status in soft tissue sarcoma; and 3) Evaluation and comparison of early, intermediate, and late fusion techniques for combining imaging and clinical data on a regression prediction task: Estimation of CT-based body composition metrics from chest radiographs. **Figure 1.1** provides an overview of the aims.

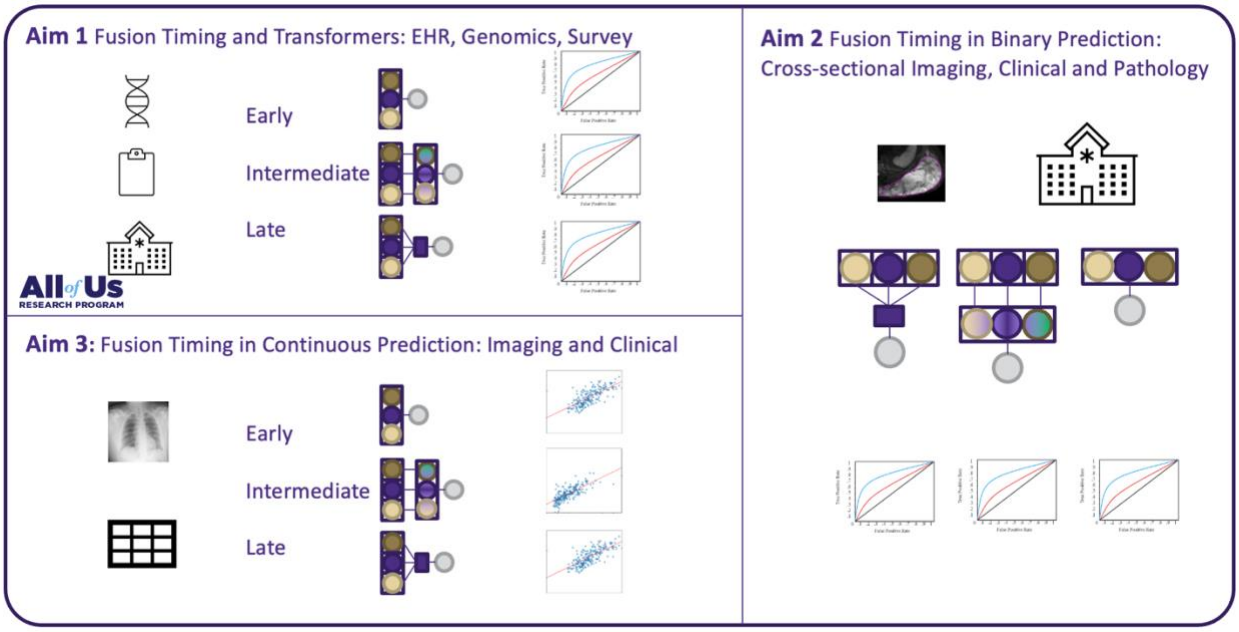

Figure 1.1 Overview of the three aims showing what type of data modalities are used and the outcomes of interest.

1.4 Outline

Starting with Chapter 2, background and significance, I will discuss the importance of multimodal predictive models in medicine and different data types that are often involved. Then, I explore the challenges in multimodal modeling, after which I introduce the concept of *data fusion* and its main variants. The later parts of chapter 2 are dedicated to description of data fusion variants and their pros and cons based on research in other domains. Finally, I will explore studies that investigated fusion in the biomedical domain and explain the gaps in the literature.

In Chapter 3, which is dedicated to Aim 1 (Evaluation and comparison of early, intermediate, and late fusion techniques for combining exposures, clinical and genomics data for disease risk prediction task using All of Us: Risk of CKD in patients with type 2 diabetes.) I developed a transformer-based, multimodal, deep learning model using data from the NIH's All of Us initiative

to predict progression to CKD in patients with T2D in a cohort of about 40,000 patients. In this aim, I showed that each of the modalities can be used in unimodal models to predict progression to CKD with the EHR based model achieving the highest performance (average AUROC of 0.73 on validation set). The use of multimodal data fusion approaches only resulted in slight improvement (average AUROC of 0.74 on validation set) in this question. This effect was only observed in the early fusion strategy and was not statistically significant.

In Chapter 4 (Aim 2, Development and assessment of the incremental value of combining a deep convolutional neural network feature extractor on imaging data and clinical data on a binary prediction task: Predict post-surgical margin status in soft tissue sarcoma), I will extend the concepts described in Chapter 1 by looking at other medical data domains and deep learning architectures. In this aim, I look at early, intermediate, and late fusion when using a convolutional neural network for longitudinal cross-sectional imaging and shallow neural network for clinical and pathology variables in a dataset of 202 patients with soft tissue sarcoma. In this aim, I demonstrated that the intermediate fusion strategy outperformed other fusion strategies in prediction of post-surgical margin status achieving an AUROC of 0.80 (0.66 – 0.95) followed by the late fusion strategy.

In Chapter 5 (Aim 3: Evaluation and comparison of early, intermediate, and late fusion techniques for combining imaging and clinical data on a regression prediction task: Estimation of CT-based body composition metrics from chest radiographs.), I will again look at fusion of imaging and clinical data, this time in the setting of multiple regression tasks to estimate body composition metrics (continuous variables) from a chest radiograph, and select clinical variables in a dataset of around 1000 cases. In this aim, I demonstrated that the late fusion approach achieved the highest

performance in predicting almost all body composition metrics closely followed by the intermediate fusion strategy.

In Chapter 6, I will provide a summary of the previous chapters and lay out the contributions made in each aim, in addition to the overarching contributions made by this body of work. I will conclude this chapter by going through limitations and future directions.

1.5 Contributions

In this dissertation, I developed three deep learning models using multimodal data fusion that can be used to 1) predict early progression to CKD in patients with T2D, 2) identify patients with higher risk for post-surgical margin involvement (with a model that achieved higher performance than any previously reported work and 3) estimate body composition metrics including subcutaneous and visceral adipose tissue volume using only a chest radiograph and four clinical variables (with the first model of its kind).

In Aim 1, I introduce a novel multimodal time and value aware transformer-based architecture using early, intermediate, and late fusion to integrate genomics, longitudinal EHR and survey responses. In this aim, I showed that early fusion can be used effectively to combine these datatypes using transformer architecture resulting in better performance on validation set and less overfitting to the training set across different fusion strategies. I also showed that late fusion can be used to combine different types of models including transformers and tree-based models (XGBoost).

In Aim 2, I showed that intermediate fusion strategy outperforms other strategies when combining imaging and clinical variables to predict post-surgical margin status. The improved performance in intermediate fusion may be attributed to cross-modal interactions between the clinical and

pathology variables (e.g., histologic subtype) and imaging features. This was followed by the late fusion strategy.

In Aim 3, I developed a multitask multimodal model to estimate CT-based body composition metrics. My results showed that across almost all metrics, the late fusion strategy outperformed other fusion approaches, closely followed by intermediate fusion. The code used to develop this model and the model weights are publicly available.

Overall, I observed that early and late fusion approaches are often easier to implement and are less likely to overfit to training data due to their lower number of trainable parameters and complexity. Early fusion may be beneficial in instances where additional feature extraction or dimensionality reduction techniques are not required, while late fusion is most beneficial when different modalities have high predictive performance. Intermediate fusion was the most challenging network to train due to its propensity to overfit. Due to that challenge, the network sections for different modalities tended to need different hyperparameters to train. I overcame this challenge by using transfer learning approaches in which networks weights for unimodal sections were initiated using pretrained unimodal network weights for each modality in intermediate fusion. A summary of lessons learned about fusion can be seen in **Figure 1.2**.

Early Fusion	Intermediate Fusion	Late Fusion
• Good choice if features can be represented with no simplification/dimensionality reduction. • Can handle missing data if appropriate architecture is used. • May lose some signal when feature extractors are used. • May help when sample size is very small for finetuning on imaging. • Performed worse even when structured data was embedded in imaging.	• May overfit to training data due to large number of trainable parameters. • Better for combination of imaging and clinical data. • Various network parts train at different speeds and may need different hyperparameters. Pretraining helps with stable training. • More design choices compared to other methods. • More costly and resource intensive.	• Can combine diverse networks like XGBoost and transformer. • Does not learn inter-modal correlations. • Less overfitting due to lower number of trainable parameters. • Performance depends on unimodal models. • Better performance when all modalities have good performance.

Figure 1.2 Lessons learned about fusion across this dissertation. Yellow: Lessons learned from Aim 1, Green: Lessons learned from Aim 2, Blue: Lessons learned from Aim 3.

1.6 Limitations

The main limitations of this dissertation include lack of breadth to provide concrete recommendations about fusion due to the complex nature of this problem. While I explored fusion of different data modality types over the three main aims, it was not possible to evaluate every combination of modalities and outcomes possible. In addition, there may be more than one approach to implement each fusion strategy (early, intermediate, and late) that may result in varying performance. Major aim-specific limitations included lack of exploration of many design choices and hyperparameters for the transformer-based model due to computational and cost constraints in Aim 1, relatively small sample sizes in Aim 2 and Aim 3, and lack of longitudinal imaging for all cases in Aim 2.

1.7 Future Directions

Given our finding that the choice of fusion can significantly influence model performance and implications with respect to computational costs and challenges, a more comprehensive evaluation of the association between various dataset and outcome attributes and the optimum choice of fusion is warranted. Ideally, such study could include a large real world or synthetic dataset in which attributes like sample size, types of modalities, the cross-modal interaction levels and outcome types can be altered.

Additionally, given the good performance of intermediate fusion for combining imaging and other variables, an exploration of the different variations and levels of intermediate fusion. Studies in other fields have shown that tailoring the depth of fusion in intermediate fusion for different data modalities may improve performance.

2. Background and Significance

In this chapter, I will explore the literature around the use of multimodal data fusion approaches in medical predictive models. First, I will explain why developing deep learning predictive models in medicine matters. Then, I will explore the role of multimodal data in improving deep learning models and the implications and challenges of using multimodal data. I will then introduce the concept of *data fusion* and its variants, and the challenges, advantages, and disadvantages of using each of these data fusion strategies based on findings in other deep learning domains. The next section will focus on studies in the biomedical domain that explored fusion techniques and the current gap in the literature. Finally, I will introduce the three aims of this dissertation in the context of the current gaps in the literature.

Disease outcome prediction is a central research focus in biomedical informatics, as it facilitates precision health related interventions and scientific discovery by enabling digital clinical trials and multiple other benefits (2). Often, deep learning approaches outperform simpler statistical or machine learning approaches due to their capability to identify non-linear complex relationships within the data (3). There are several informatics challenges that must be addressed to enable effective development of predictive models. First, biomedical data consists of different data modalities such as medical imaging and clinical notes that are often incompatible in their preprocessed formats, and the computational models and pipelines typically used to extract insights from these data types are specialized and non-interoperable (4). Additionally, data are frequently siloed across institutions and cannot be easily aggregated due to restrictive data-sharing policies. Compounding these issues is the lack of standardized ontologies, terminologies, and annotations across datasets, which hampers cross-institutional analyses (2). These challenges collectively complicate the integration and analysis of biomedical data to develop predictive

medical models. In this dissertation, I will focus on challenges related to integration of multimodal biomedical data in deep learning models.

Multimodal medical data has the potential to improve medical deep learning models, as different data modalities often provide complementary information about the patient (4,5). Multimodal models are closer to actual medical practice as a physician usually consider data from multiple sources when making a decision about the patient. Many studies have demonstrated the advantage of multimodal modeling over unimodal models (4,6,7). The process of combining multiple data modalities in machine learning is called *data fusion* (8). Three main approaches of data fusion have been introduced in literature. These include early fusion (feature level), intermediate fusion, and late fusion (decision level) (4,8,9). Few studies have compared the influence of the choice of data fusion on final model performance in medicine. Currently, literature recommends experimenting with all fusion strategies when developing deep learning models (4). However, most studies apply one data fusion strategy without sufficient justification for why that strategy was selected (4,10). Over the next three chapters. I will evaluate the performance of the three main data fusion strategies across a series of three medically relevant deep learning models which differ with respect to sample size, data modalities being combined, network architectures and outcome types including categorical and continuous outcomes.

Multimodal deep learning models are increasing in popularity in biomedicine (8). This is because each modality contains both marginal (modality specific) and joint (cross-modality) information that can be extracted if other relevant modalities are available. However, many challenges need to be overcome before data from different modalities can be brought together in a deep learning model (4,11,12). For context, it is helpful to first review some commonly used types of medical data that can be used in predictive modeling. Each of these data types have unique characteristics

and analysis requirements that make multimodal data fusion more challenging (12). These include multiple categories including the electronic health records (EHR) data, which in itself can be divided into multiple categories including structured and unstructured EHR data. Structured EHR data can be further categorized semantically into demographics, diagnosis codes, measurements and laboratory tests, procedures, and medications. Unstructured EHR data include various types of notes and reports and are usually text-based (13). Imaging is another large category of medical data that includes 2D or cross-sectional radiologic images (CT scan, MRI, PET), functional images (fMRI, echocardiography) and pathology images that depict cellular level details. Another category of medical data includes the various types of omics data including genomics, transcriptomics, proteomics and epigenomic data. The next category of medical data includes survey data that are information that are collected from individuals in form of interviews or questionnaires and can include information about their lifestyle, personal and family medical history, and social determinants of health. Recently, wearables have introduced a new type of medical data that includes continuous monitoring of variables like heart rate, oxygen saturation and sleep (14).

Integration of these data modalities poses specific challenges. Biomedical data is diverse with respect to size of data points, dimensionality, missingness amount and analytic methods (12). For example, whole genome sequence data could be as large as 200 Gigabytes, while imaging data can range from a few thousand pixels in a chest x-ray to billions of pixels in a pathology slide. In addition, medical imaging can be two dimensional or cross-sectional, and functional imaging techniques have the added dimension of time as well. These differences in data characteristics have resulted in differences in analytic methods used for any of these data modalities (12).

Traditionally, researchers have resorted to manual feature extraction strategies to incorporate EHR data into their models (13). Commonly a list of select clinical variables known to be relevant to the question at hand are extracted. These variables are subsequently fed into algorithms ranging from simple linear regression to more sophisticated deep learning or tree-based methods like XGBoost. Knowledge injection, in the form of selecting and curating relevant clinical variables happens in this process. However, feature extraction from the EHR data may introduce limitations. Often these curated variables do not encode longitudinal nature of the clinical data. For example, a predictive model may use the average HbA1C value over the last year to predict adverse outcomes in diabetes. However, such model won't take into consideration the trend of change in HbA1C measurements over that period. Additionally, using the traditional approaches, researcher need to identify methods to address data missingness (a common problem in the EHRs). These may include imputation strategies or removal of patients who do not have all the necessary data. Both of these approaches risk introducing unwanted bias into the research. Newer approaches in EHR data analysis try to mitigate these limitations by incorporating more information from the EHRs (13). Deep learning approaches that are capable of analyzing timeseries data like recurrent neural networks (RNNs) and, more recently, transformers, can accept sequences of EHR events (13). BEHRT which is one of the earlier works in this domain, accepts a sequence of all diagnosis codes recorded in the EHR for the patient, taking into account their relative position in the sequence (15). More recent architectures take this further by utilizing other information stored in the EHR, including laboratory values and medications, and by introducing time difference and position embeddings (16).

Imaging data often require specialized feature extraction techniques or deep learning architectures due to their large size and dimensionality (10). Like EHR data, earlier efforts to incorporate

imaging data included utilizing feature extraction strategies to convert an image to a list of predefined variables. This can be done either using expert evaluations (a list of radiologic features extracted by a radiologist), or quantitative imaging approaches, including radiomic feature extraction(17). However, imaging data are rich in details, and these strategies may inadvertently result in loss of important signals that are not encoded by feature extraction strategies. The advent of deep learning strategies has introduced deep learning architectures like convolutional neural networks and vision transformers (18) that are specifically designed to learn representations from 2D and 3D imaging data.

Omics data is characterized by high dimensionality, sparsity and missingness (12). Whole genome sequence data can contain hundreds of millions of single nucleotide polymorphisms (SNPs) (19). Hence, utilization of knowledge injection and feature extraction strategies are necessary for modeling to reduce the dimensionality of these data. Common approaches include using genome wide association studies (GWAS) to generate and calculate *polygenic risk scores* (PRS) (20,21). These are trait-specific aggregate risk scores that are calculated by accumulating the effect sizes of relevant SNPs. Other approaches can include identification of relevant SNPs from previous literature.

As described earlier, three main approaches for combination of different data modalities are described in the literature that include early, intermediate, and late fusion. Each of these approaches has been studied extensively in other domains of machine learning (22). However, few studies have compared the implication of these fusion strategies in biomedical use cases (4). In the following paragraphs, I will explain each of these fusion strategies and provide examples of their application in literature.

Early fusion involves the combination of data from all modalities before any learning takes place. This can be done either by direct concatenation of raw values for all input modalities or by using feature extraction strategies (e.g. radiomics, PRS, …), with concatenation of resulting variables (8). The latter is specifically useful when input data modalities have large differences in dimension or size. For example, if one tries to combine cross-sectional 3D MR images with a list of clinical variables without using any feature extraction strategy, the signal from the clinical variables may easily be lost when combined with the millions of voxel values in the imaging data. The simplicity of early fusion has made it the most popular data fusion strategy used in biomedical literature (4,10). However, early fusion has weaknesses. Studies have shown that early combination of data modalities may prevent the model from learning intra-modality features very well. In fact, while early fusion may perform well when variables from different modalities have fine grained relationships with each other (22). On the other hand, early fusion may suffer when complex features from each modality are related to each other as the model has not had the time to learn those complex features (8,23). In addition, the use of feature extraction strategies may result in oversimplification and loss of signal in some modalities (8). Finally, if only raw variables are concatenated together, early fusion may become computationally expensive as the number of input variables may become too high.

Intermediate fusion involves the use of modality specific networks to process each modality so that embeddings are generated from each modality that can be combined together and further processed by the rest of the network (8). In intermediate fusion, the whole network, including the modality specific parts and the multimodal section train together, enabling the network to learn rich modality specific embeddings that are influenced by other modalities as well. Multiple choices exists about after what amount of modality specific processing the modalities should be fused,

what fusion methods should be used (concatenation, cross-modal attention, …) and what types of networks should be used for each modality (8). However, the complexity of intermediate fusion networks makes them harder to optimize and more likely to overfit if a necessary sample size is unavailable.

Late fusion involves the training of modality specific networks and integrating the final networks at decision level, meaning that once each modality specific network makes a prediction, all predictions are aggregated together to come up with the final prediction (8). Various approaches like majority-voting or averaging are used to make the final prediction. More advanced methods of combining prediction s could include the use of a small fully-connected neural network (meta-learning). Late fusion enables the model to fully learn complex intra-modal features relevant to the task at hand with less probability of overfitting due to the smaller number of trainable parameters and less complexity in the network architecture. This can be particularly useful in cases that the sample size is not large enough for the model to learn joint representations between the modalities. However, inter-modality interactions are not captured but at the very end. This would hinder the model's ability to learn more fine-grained interactions between variables in different modalities (8).

Few studies have evaluated and compared the performance of early, intermediate, and late fusion strategies in the biomedical domain. Notably, Huang et al. compared early, intermediate, and late fusion in identifying patients with pulmonary embolism when comparing imaging and clinical data. They found that late fusion outperformed other strategies achieving an AUROC of 0.947 (0.946 – 0.948) (10). In another study by Roest et al. only early and late fusion strategies in detecting prostate cancer from MR images and clinical variables. They found that early fusion achieved slightly higher performance. However, they use suspicion levels generated from another

deep learning model as a feature in both the early fusion and late fusion models. Hence, It could be argued that the early fusion model was not a true early fusion model (7). A study by Ying An et al. combined longitudinal diagnosis codes, laboratory tests and medication data from the EHR using the transformer architecture. They found that intermediate fusion outperformed early fusion in prediction of clinical endpoints in ICU patients (24). However, they did not compare these models with late fusion models.

Given the utility of using multiple data domains in developing predictive models in medicine, and the existence of various fusion strategies, my overarching question is evaluating multi-modal data fusion approaches including early fusion, intermediate fusion, and late fusion strategies in developing deep learning predictive models when combining multiple data domains. To answer this question, I explore three separate biomedically relevant questions that benefit from multimodal fusion. Each of these three questions is distinct in the types of biomedical data that are included, the types of deep learning models that are employed or the outcome measure type. My goal is that by looking at data fusion across these diverse set of examples, I shed light on the advantages and disadvantages of each of these fusion strategies and highlight some of the challenges in implementing each of them. In my first aim, I explore the gap in literature regarding fusion of clinical EHR data, genomic data and lifestyle survey data in predicting progression to chronic kidney disease in patients with type 2 diabetes (binary outcome prediction). A novel state of the art transformer architecture is used to analyze the EHR data. In the second aim, I look at the fusion of volumetric imaging data with clinical and pathologic data to predict involvement of post-surgical tissue margin in patients with soft tissue sarcoma (binary outcome prediction) and address the gap regarding the utility of fusion approaches in combining cross-sectional imaging and clinical variables. In the third aim, I explore the fusion of 2D imaging and clinical data to predict

a series of body composition metrics to address the gap in literature regarding the application of fusion strategies to predict continuous outcomes when combining imaging and clinical data.

In the next chapter, I will dive into more details about aim 1 and the fusion strategies used to combine the EHR, genomics and survey data.

3. Chapter 3: Aim 1, Evaluation and comparison of early, intermediate, and late fusion techniques for combining exposures, clinical and genomics data for disease risk prediction task using All of Us: Risk of CKD in patients with type 2 diabetes.

3.1 Introduction

As described in the background and significance (Chapter 2), the overarching question of my dissertation is "evaluating multi-modal data fusion approaches including early fusion, intermediate fusion, and late fusion strategies in developing deep learning predictive models when combining multiple data domains". The first aim is "Evaluation and comparison of early, intermediate, and late fusion techniques for combining exposures, clinical and genomics data for disease risk prediction task using All of Us: Risk of CKD in patients with type 2 diabetes.". This aim addresses the gap in the literature regarding the utility of early, intermediate, and late fusion when combining EHR, genomic and survey data, using the transformer architecture. This will be the first step towards answering the overarching question. To address this gap, we will explore fusion of clinical EHR data analyzed with genomics and survey data extracted from the NIH All of Us dataset while using the transformer deep learning architecture to predict progression to chronic kidney disease in patients with type 2 diabetes. The transformer architecture provides benefits that make it a strong candidate for analysis of longitudinal electronic health record (EHR) data, which I will discuss further in the next section. Transformers have been previously used to analyze EHR data. However, those implementations usually had limitations, which will be discussed in this chapter. This aim consists of two main parts. First, I will develop a transformer architecture to analyze longitudinal EHR data that takes into consideration the sequence of events, the time difference between the recorded events, and the value assigned to the events where applicable (e.g. laboratory test results). Then, using early, intermediate, and late fusion strategies, I will integrate the EHR data with

structured data extracted from survey responses related to family history and social determinants of health, and domain specific genomic data to predict progression to chronic kidney disease (CKD) in patients with type 2 diabetes (T2D). Using this strategy, I will first demonstrate the ability of EHR data, genomics data and survey data to predict progression to CKD in patients with T2D separately. Then, I will show how combining these complementary data modalities affects final model performance. Lastly, I will compare early, intermediate, and late fusion in this scenario to determine the best fusion strategy given this prediction problem. For this chapter, I will be using the NIH's All of Us dataset (1).

3.2 Background

Type 2 diabetes, one of the most prevalent chronic non-communicable diseases, predisposes patients to several debilitating adverse outcomes, including cardiovascular disease, stroke, peripheral neuropathy, and kidney disease. Chronic kidney disease, which is referred to as diabetic kidney disease when the other underlying causes are considered unlikely, is one of the most detrimental adverse complications of T2D. Chronic kidney disease is defined as irreversible decline in kidney function as measured by the glomerular filtration rate and proteinuria. Studies have shown that approximately 50% of patients with T2D eventually progress to CKD (25,26). Many of these patients will eventually need dialysis or kidney transplantation. It is possible to slow the decline of kidney function using various strategies, including maintaining a tight glycemic control and using angiotensin converting enzyme inhibitors (ACE inhibitors) or angiotensin receptor blockers (ARB). Studies have shown that multifactorial interventions that include tight glycemic control, lowering low density lipoprotein and maintaining blood pressure control levels below 120/75 can significantly reduce kidney function deterioration in patients with type 2

diabetes and even prevent CKD (27,28). However, these strategies need to be considered carefully to balance benefits and adverse effects like hypoglycemia (27).

Multiple studies have investigated developing risk models for progression to chronic kidney disease in patients with T2D. Most studies have focused on using pre-selected clinical features, including relevant comorbidities and laboratory test values, such as HbA1C and baseline creatinine, to predict progression. Separately, many studies have investigated the genetic bases of chronic kidney disease and specifically diabetic kidney disease. Notably, a study performed on the data from the Chronic Kidney Disease Prognosis Consortium (CKD-PC) showed that a model using variables like demographic variables (age, sex, race/ethnicity), estimated glomerular filtration rate (eGFR), history of cardiovascular disease, smoking status, hypertension, body mass index (BMI), and albuminuria, diabetes medication use, and hemoglobin A1c levels achieved a Harrel's C-score of 0.80 in predicting progression to CKD in patients with T2D within 5 years (29). This risk score was trained on 15 cohorts of patients with T2D and was externally validated on multiple other cohorts.

Machine learning approaches have also been used to predict DKD. A study performed in China used data from 73,101 patients with type 2 diabetes to identify those with DKD. This study used as many as 60 clinical variables extracted from the EHR for this predictive task. They achieved an AUROC of 0.97 using the CatBoost algorithm (30). However, the outcome definition in this study was unclear and it was not apparent whether any of the observations from the EHR had occurred after the patient progressed to DKD which would indicate the model could classify patients correctly, but not necessarily be valid for prediction. Additionally, their most important predictive feature was existence of non-DKD CKD. In many studies patients who already have CKD are excluded from the analysis.

A meta-analysis by Li et al. found that out of all studies that used various machine learning approaches to predict DKD in T2D, the average AUROC was 0.84 (95% CI 0.79 – 0.89) with deep learning approaches achieving higher values (AUROC of 0.86) (31). However, the studies were diverse with respect to the study design, the exact definition of the outcome measure and other relevant factors.

Other data modalities have been used to predict progression to DKD as well. A recent study used a combination 329 variables extracted from the EHR, lifestyle information, retinal imaging, genetics and blood metabolomics in a population of 1365 Chinese, Malay and Indian people with T2D. This study only used early fusion to predict the occurrence of DKD in patients with T2D in a 6-year follow up. They used a semiautomated computer program to extract predefined retinal imaging parameters. They used a set of 76 important single nucleotide polymorphisms (SNPs) based on previous studies for their genomic data (32). These variants were not specific to diabetic kidney disease. This study achieved an AUROC of 0.85(0.84 – 0.86).

Additionally, multiple studies have investigated the genetic risk factors for DKD. However, these associations while significant, are usually not very strong (33–36).

While traditional machine learning and deep learning approaches have been quite successful in many biomedical prediction tasks, they have limitations that may negatively influence their performance or limit their large-scale application. One important limitation of traditional approaches is the requirement for the developers to curate a set of relevant features for input to these models. For example, in the case of prediction of DKD in T2D, scientists often have to use computational phenotyping approaches, including curation of a list of relevant variables and their definitions to be extracted from the data. These may include the mean fasting blood glucose level over the past year, the value of the most recent HbA1c laboratory test, and history of relevant

comorbidities. Using this approach, the model may lose many of the important details available in EHR observations, including the longitudinal relationships between the various events and laboratory test values, the important clinical events that were not considered relevant by the scientists at the time they curated their list of input variables, and the fine-grained trajectories of each of concept recorded in the EHR (e.g. changes in HbA1c measures over time, which would be lost with if taking the mean HbA1c measurements over a specified time period.).

Given these limitations, some researchers have attempted to use the power of transformers, a deep learning architecture often used in natural language processing (NLP) and computer vision, which can analyze long sequences of data and extract longitudinal relationships (37,38). Earlier examples of use of transformers include models like BEHRT (39), HI-BEHRT(40), time-aware transformer-based hierarchical attention network (TERTIAN)(24) and the more recently developed hybrid value-aware transformer (HVAT)(16). While earlier versions like BEHRT only ingested sequences of diagnosis codes, future versions included time difference encoding (TERTIAN) and value encoding (HVAT) as well. Studies have shown that pretraining the transformer models using strategies like masked language modeling (albeit altered to better fit the EHR data) may improve the performance of these transformer models.

There is a paucity of studies exploring the fusion of transformer models with other deep learning architectures and data types in medicine. One example is the TERTIAN model that explored early, intermediate, and late fusion of the various EHR data subtypes including diagnosis codes, lab values and medication prescriptions. Their study found that a form of intermediate fusion of these data yields the best performance (24).

Another notable work in this field is done by Zhou et al. that introduce a multimodal transformer model that utilizes a version of intermediate fusion with self and cross modal attention layers to

fuse structured EHR data, text data and imaging data. The authors of this paper demonstrated that this approach improves the performance of their model in prediction of adverse events in patients with COVID-19 compared to early and late fusion approaches (41). This work did not focus on the longitudinal nature of EHR data and used structured data extracted from the EHR for their EHR section. In addition, based on the limitations mentioned in their work, their model cannot not handle missing data.

We identified multiple gaps in the literature when it comes to prediction of DKD in T2D using machine learning. First, no study has attempted to predict DKD by combining genomic, survey and EHR data. Works that have integrated genomics, have used only a limited number of SNPs. Additionally, all studies identified have relied on pre-selected variables that may limit the model's ability to use longitudinal nature of clinical data. We did not find any study that used longitudinal EHR data using RNNs or transformers to predict DKD. Finally, we did not find a study that focuses on early detection of patients at high risk for DKD as most studies included patients that may have had T2D for years prior to their index date.

In this aim, we will address these gaps by employing a introducing a multimodal transformer-based deep learning architecture that is designed to integrate genomic, EHR and survey response data from All of Us to predict progression to CKD in patients with T2D. The model will utilize longitudinal EHR data, CKD related polygenic risk scores, pathogenic and likely pathogenic variants from ClinVar, variables extracted from survey responses. To address the gap related to lack of exploration of fusion strategies between genomic, EHR and survey data, we will introduce three variants of our model, based on early, intermediate, and late fusion strategies.

3.3 Significance

As discussed in the previous section, most efforts to predict progression to CKD in patients with T2D involve studies performed in unimodal data. In addition, most studies have relied on a manually curated list of clinical variables. These limitations may limit the application of existing models to patients who have all those clinical variables available.

Motivated by the goal of developing models that utilize multiple data modalities, consider longitudinal relationships of events, can handle missing data automatically and do not rely on domain specific input variable curation; I propose a multimodal transformer-based deep learning architecture to generate embeddings from the patient EHR records, relevant genomic data and survey information about their family history, lifestyle and social determinants of health to predict the 5 year risk of progressing to CKD in patients with T2D very early after the initial diagnosis of T2D. This model will use all data prior to the diagnosis of T2D, in addition to the EHR data over the first year after the diagnosis of T2D. This early screening strategy will enable the physicians to identify high risk individuals and implement preventive measures earlier in the course of the disease.

In the course of developing this model, I will explore the three main data fusion approaches (early, intermediate and late fusion) in the context of combing EHR data analyzed by a transformer model with genomic data and structured survey data. This aim will address the current gap in the literature by being the first work to combine these three data types using the transformer architecture and being the first model to utilize all these data types under various data fusion approaches to predict progression to CKD in patients with T2D. To the best of my knowledge, this is the first work that explores fusion of these three data types using a transformer model. Our approach can be used in the future to combine genomic, survey and EHR data from All of Us and other similar datasets for

other use cases as well. In addition, it will provide a comprehensive screening tool for patients with T2D to identify those at a high risk for progression to CKD early in their disease course. In the next section, we will go over the technical details on how this aim was implemented.

3.4 Materials and Methods

3.4.1 Data Sources and Study Population

This study was performed on data from the NIH's All of Us initiative (1). The All of Us dataset contains deidentified information for a diverse cohort of volunteers. This includes EHR data for most subject and whole genome sequencing (WGS) data for many subjects as well. Multiple surveys are available for many patients that detail their personal history, family history, lifestyle, social determinants of health and other health-related information.

We identified individuals with type 2 diabetes by searching the EHR data for type 2 diabetes diagnosis codes extracted using the All of Us cohort builder. Additionally, among these individuals, we identified those with chronic kidney disease using diagnosis codes extracted the same way. The date of the first instance of appearance of any T2D-related or CKD-related concept code was considered to be the initial date of diagnosis for T2D and CKD, respectively. Subjects for whom the initial CKD diagnosis date was before or within one year following the initial diagnosis of T2D were excluded from the analysis.

3.4.2 Patient Timeline Generation

For each patient, using EHR diagnosis codes, measurements, laboratory test values and medications, we generated a timeline with concept codes for each event in their medical record placed sequentially, starting from the first event recorded in the EHR. Events up to one year after the initial diagnosis of T2D were included because, for many patients, additional work up is done

after the diagnosis of T2D to both determine adverse outcomes that have already occurred and establish a baseline for future assessments. The timeline included three components for each entry. The first component was the concept code attributed in the All of Us data to that entry. These codes are unique for each type of event and there exists a mapping between these and the relevant SNOMED , LOINC(42) , and other ontologies. The second component of each entry was the value attributed to that entry. For laboratory tests and other measurements, the value was assigned as the numerical result of the measurement. For diagnosis codes and medications, we passed the number 1 as an indication that this diagnosis code was recorded, or the medication prescribed (in this case, a value 0 means that this diagnosis code was not recorded in the EHR). The third component of each entry was the time difference between the entry and the date of birth in months. This component allows the model to exactly identify when this entry happened.

3.4.3 Demographic Information

Three relevant demographic variables were extracted an included at the beginning of the patient's timeline including the date of birth, sex at birth and gender.

3.4.4 Genomic Data

Due to the large volume of whole genome sequencing data, we implemented feature selection strategies and knowledge injection from the literature to identify the most important genomic regions for progression to CKD. We prepared two sets of relevant genomic features. First, we searched for polygenic risk scores (PGS) for CKD in literature developed or validated on populations with diverse ancestry (as is the case with All of Us). Additionally, we used the ClinVar dataset to identify all the pathogenic and likely pathogenic variants associated with CKD and generated a dataset with the number of alternate alleles for each variant for each patient.

To identify relevant PGSs, we searched the pgscatalog website (43,44) which is a repository hosting a curated list of PGSs with information regarding their development population and validation populations. At the time of our search in September 2023, although, all the available PGSs were developed on cohorts with European ancestry, we identified two PGSs that were validated on a diverse cohort of patients with ancestries from Europe, Africa, East Asia and South Asia. One PGS with 158 single nucleotide polymorphisms (SNPs) (45) and one with 41,419 SNPs with none-zero effect-size (36). The ACAF (allele count, allele frequency) variant call table was used which is a table filtered by allele frequency (AF > 1%) and allele count (AC > 100) in each ancestry subpopulation. I used LiftOver (version 1.3.2) (46) to convert any SNP location in other reference genomes to the default All of Us reference genome GRCh38. The Hail package version 0.2.130 was used for all analysis. The PGS scoring files were harmonized with the All of Us genotype data by matching the variant locations based on GRCh38 reference genome and aligning the alleles. The variants with allele mismatch or low allele frequency or count were excluded as they were not available in the ACAF variant call table. The final PGS was calculated per subject as the sum of the effect size of each allele multiplied by the number of effect alleles in that location. A second dataset was created by querying the ClinVar database (accessed on February 10th 2025) (47) to identify pathogenic and likely pathogenic variants for phenotypes related to CKD. We intersected the resulting list with the ACAF table from the All of Us Research Program using the GRCh38 location excluding variants with low allele counts, low allele frequencies or inconsistent alleles. The dataset, for each patients recorded the number of pathogenic or likely pathogenic alleles each patient had for each genomic location.

3.4.5 Survey Data

Information extracted from three surveys were included in the study. These surveys included the family history section of the personal and family history survey, the social determinant of health survey and the lifestyle survey. The personal history section of the personal and family history survey was not included as this survey may have been administered after our index date of the initial diagnosis of T2D.

For family history, if for any disease the response of the patient indicated that any member of their family (including parents, grandparents, and siblings) had a disease, the response will be recorded as positive. All questions from the survey were included. For questions in which the volunteer skipped or did not response, missing response was recorded.

For the lifestyle survey, the omop2survey (version 0.0.40) package was used to collect the responses and convert them to OMOP data model compatible variables. The questions in the lifestyle survey included information about alcohol, tobacco and recreational/prescription drug consumption including the duration, quantity, and frequency.

For the SDOH survey, we used the methods described in literature and recommended by All of Us to score the responses (48,49). The scores calculated are the Physical Activity Neighborhood Environment Scale (PANES) walking and bicycling score, the PANES crime and safety score (50), loneliness, social support, social cohesion, everyday discrimination, discrimination in medical setting, perceived stress, daily spiritual experience, neighborhood social disorder, food insecurity, housing instability and housing quality with higher scores indicating a better score.

3.4.6 Outcome Definition

We defined a binary outcome of whether a patient will get a diagnosis of chronic kidney disease between years 1 and 5 after their initial diagnosis of T2D. To do this, we searched for diagnosis

codes related to CKD in the patient's EHR in the aforementioned timeframe. The All of Us concept codes used for T2D, and CKD can be seen in **Table 3.1**. These concept codes and any children concept codes were used for establishing the diagnosis of T2D and CKD.

Table 3.1. All of Us concept codes used to identify patients with type 2 diabetes (T2D) and chronic kidney disease (CKD). The first instance of the occurrence of any of these concept codes or their children codes was recorded as the initial diagnosis date.

Condition	All of Us Concept Codes
Type 2 Diabetes	201530, 201826, 376065, 443729, 443731, 443732, 443733, 443734, 4063043, 4099216, 4099651, 4129519, 4130162, 4140466, 4193704, 4196141, 4200875, 4215719, 4221495, 4222415, 4222876, 4226121, 4228443, 4230254, 4304377, 35626070, 36712686, 36712687, 36714116, 37016349, 37016354, 37016768, 37017432, 43530656, 43530685, 43530689, 43530690, 43531010, 43531562, 43531563, 43531564, 43531578, 43531616, 43531651, 43531653, 45757363, 45757435, 45757449, 45757474, 45757499, 45770830, 45770881, 45773064
Chronic Kidney Disease	443597, 443601, 443611, 443612, 443614, 43531578, 44784621, 44782690, 44784638, 45763854, 45763855, 46271022

3.4.7 Modeling Strategy and Fusion Approaches

The modeling strategy used varied by the type of fusion strategy to conform to the requirements of that strategy. For the sequential EHR data, the RoBERTa transformer architecture was used as the base transformer model (51). The transformer was altered to accept additional time difference and value embeddings which were summed up with the original embedding the roBERTa model generated for each token. The final hidden layer of the network was used for the masked concept modeling task while the output of the model for the [CLS] token was used for the binary classification task using a fully connected layer and a sigmoid activation function to generate probabilities between 0 and 1. The maximum model input length was defined to be 514 tokens at each time.

Before the sequence can be passed to the transformer model, some preprocessing steps are required to be performed to transform the sequence of events into an acceptable format for the model (a sequence of integers). First, a vocabulary is generated using all the concept codes in the EHR to

link each code into a numerical representation. Then, a tokenizer is defined, which takes in the string sequence and, using the vocabulary, converts it into a sequence of integers and adds any additional tokens necessary for the model. Finally, before passing the input, each batch of sequences is passed through a data collator to ensure the sequence is padded or truncated as necessary to conform to the length constraints of the model. The data collator takes care of masking as needed. These components were altered from their original form to account for the differences between the EHR based inputs and the traditional string-based text input. The specific details of each component are discussed below.

3.4.7.1 Vocabulary Definition

The vocabulary of the model was defined as a list of all the concept codes that appeared in the All of Us EHR data sorted in an alphabetic order and each token was assigned a unique integer value starting from 0 and increasing in the increments of 1. Additionally, tokens for masking [MASK], start of the sequence [CLS], end of sequence [END], padding [PAD] and unknown tokens [UNK] were added to the vocabulary. For early fusion, the questions in the surveys and the SNP locations, and a token for each PRS was added to the vocabulary as well as these were also incorporated to the sequence.

3.4.7.2 Tokenizer Definition

A custom tokenizer was built using the default transformers tokenizer was altered to handle the time difference and value embeddings. The tokenizer was configured to not do any truncation or padding at this stage as these were done by the data collator. The tokenizer performed the following tasks. First, it replaced each token with its corresponding integer value from the vocabulary. Additionally, the tokenizer added the beginning of sequences [CLS] and the end of sequence [END] tokens to the beginning and the end of the sequence. A time difference of 0 and a value of

1 was arbitrary assigned to these new tokens. Time difference and value tokens were multiplied by 1e-4 to prevent exploding gradients or loss.

3.4.7.3 Data Collator Definition

A custom data collator was built based on the default transformer data collator to consistent truncation and padding on the input sequences. The data collator padded all input sequences in each batch to match the size of the longest sequence in the batch (or the maximum input length). Since the maximum input length of the model was 514 tokens, each sequence that contained more tokens was truncated. For such sequences, we randomly omitted concept codes and their corresponding time difference value and numerical values from the sequences using an incremental probability distribution to favor the retainment of events closer to the end of the timeline. The probability that each token would be retained was calculated using the following formula in which the increase factor was set to 1.005 and P_i is the probability that token i is retained:

$$P_i = \frac{increase\_factor * token\_position_i}{\sum_{j=1}^{sequence\_length} increase\_factor * token\_position_j}$$

For the masked concept modeling step, the data collator randomly masked 20% of input tokens, replacing them with the [mask] token value. 20% of time differences and 20% of values were also masked independently.

3.4.7.4 Token Embedding Generation

Transformer models usually generate position embeddings to encode the location of each token within the sequence. In our modified transformer, in addition to position embeddings, we generate time difference embeddings and values embeddings that are added to the original token and

position embeddings using a summing function. The embeddings are learned using a shallow neural network for each of these features during the training. The final token embedding is calculated using the following formula:

$$E_{Token} = E_{Concept\ code} + E_{Position} + E_{Time\ Difference} + E_{Value}$$

3.4.7.5 Masked Concept Modeling

For the pre-training step, we employed a strategy like the masked language modeling concept used in training language models. Our strategy is distinct in that we not only mask the tokens, but also independently mask the values for time difference and values as well. During the pretraining steps, the model will predict each of the missing inputs separately using the last hidden layer of the network. For the input tokens, the cross-entropy loss is used. For the time difference and value predictions, we use Huber loss, as these are continuous variables. Huber loss is a combined loss consisting of L1 loss for large differences and L2 loss for the smaller differences in values.

3.4.7.6 Genomic Model

The genomic model consisted of a XGBoost model trained using five-fold cross-validation on the combination of the train and validation sets used to train the multimodal models. Hyperparameters that were tuned included the number of estimators, maximum depth, minimum child weight and alpha and lambda regularization.

The genomic part of the intermediate fusion model consisted of a shallow neural network with three fully connected layers with intermediate hidden layer sizes of 256 and 64 and one classification head with sigmoid activation function. In addition, this model utilized two random dropout layers with a probability of 50%. The model used the two PRSs along with the ClinVar

dataset to predict progression to CKD within the next 5 years. The binary cross-entropy loss was used to train this network.

3.4.7.7 Survey Model

The unimodal survey model was also an XGBoost model trained using the same training strategy as the genomic data.

The survey model for the intermediate fusion strategy consisted of shallow neural network with two fully connected layer and hidden layer size of 64 and one classification head with sigmoid activation function. This model had one random dropout layer with a probability of 50%. It was trained using the binary cross-entropy loss. The last hidden state of both the genomic and the survey model were used as embeddings for the intermediate fusion strategy discussed later. The predictions of these two networks were used for the late fusion network.

3.4.7.8 Fusion Strategies

3.4.7.8.1 Early Fusion

For early fusion, I incorporated the genomic and survey results into the EHR sequence by adding each non-zero variable as a token to the beginning of the sequence with the time difference set to zero and the value set to match the actual value of the variable. For example, for each genomic location in the ClinVar dataset, if the patient had one or two effect alleles, the token for that locus was included in the sequence and the value would be 1 or 2 depending on the number of effect alleles. The resulting sequence was passed to the rest of the transformer model for the binary classification task.

3.4.7.8.2 Intermediate Fusion

The intermediate fusion network consists of a unified architecture that combines the embeddings generated from the EHR data using the transformer model and the last hidden states of the genomic and survey data. Three multi-head cross-attention layers were used to concatenate the embeddings into a vector that was fed into the final fully connected layers with a single classification head with sigmoid activation function. The pretrained weights from each of the unimodal models were initially loaded into each component of the network. Then the model was fine-tuned using all modalities.

3.4.7.8.3 Late Fusion

For the late fusion strategy, the predictions of the unimodal EHR, genomic and survey were aggregated into a vector with the size of three. Since the late fusion strategy allows for integration of different types of models, the unimodal XGBoost based models were used instead of their shallow neural network counterparts due to slightly higher performance. A shallow neural network with hidden size of 10 and a single classification head was used to make the final prediction using the probabilities provided by each of the models. **Figure 3.1** provides a graphical representation

of each of these fusion strategies and their input data.

Figure 3.1. Overview of the architectures of early, intermediate, and late fusion strategies.

3.4.7.9 Loss Function

The binary cross-entropy loss was used for the binary classification task.

3.4.7.10 Handling of Missing Data

For the data from the EHR, the only requirement was for each of the cases to have some sort of EHR record. We did not filter for any specific variable or encounter. This strategy allows the model to inherently be able to handle cases that are missing relevant variables like HbA1c measurements, blood pressure, etc.

For the structured data that were extracted from the surveys and the ClinVar genomic table, all missing variables were replaced with a constant negative number (-1) to indicate their missingness.

For the genomic unimodal model, only cases that had whole genomic sequencing data were included during training. For the survey-only models, patients were included if they had the family history and lifestyle surveys available. Any missing responses or any missing fields within the genomic data were replaced by -1.

3.4.7.11 Model Evaluation

Multiple outcome metrics, including area under the receiver operator curve (AUROC), area under the precision recall curve (AUPRC), sensitivity, specificity, positive predictive value and negative predictive value were reported over the training, validation and test sets. The probability threshold for calculation of sensitivity, specificity and other threshold dependent metrics was calculated using the Youden Index.

3.4.7.12 Fairness Evaluation

We calculated and compared the above performance metrics within the various subgroups of our data. The different subgroups included different age groups, sex, race and ethnicity and genetic ancestry.

3.4.7.13 Model Selection and Hyperparameter Tuning

We split the dataset into training, validation, and testing sets with a ratio of 0.7, 0.1, 0.2. We performed the split on the patient level and the split was consistent across experiments. Random search over the hyperparameter space was performed to identify the best performing model over the validation set. The best performing models' performance on the hold-out testing set was reported as the final model performance.

3.4.8 Statistical Analysis

A significance level threshold of $p < 0.05$ was applied to all statistical analyses. The DeLong method was employed to compute 95% confidence intervals for the area under the receiver operating characteristic curve (AUROC) and to perform pairwise comparisons of model performance. To assess differences across data splits, analysis of variance (ANOVA) was used for continuous variables, while the chi-square test of independence was applied to categorical variables. All statistical tests were two-sided unless otherwise specified. All analysis was performed on a google cloud platform instance with a NVIDIA Tesla T4 GPU. The packages used included PyTorch, Transformers, Datasets, Hail, Confidenceinterval, sci-kit learn, matplotlib were used.

3.5 Results

3.5.1 Cohort Characteristics

The final cohort consisted of 46,961 individuals with the diagnosis of T2D who did not have chronic kidney disease at the time of the initial T2D diagnosis. Out of these, 30,065 were randomly assigned to the training set, 7,502 were assigned to the validation set, and 9,394 to the testing set. The male to female ratio was 0.42. The average age of patients at the time of diagnosis of T2D was 54 (±13) years old. Out of all patients, 4380 (9%) were diagnosed with CKD within the next 6 years after their diagnosis of T2D (excluding the first year). **Table 3.2** summarizes the cohort characteristics, and the cohort flowchart is available in **Figure 3.2**.

Table 3.2. This table summarizes the cohort characteristics among the train, validation, and testing cohorts. We used statistical tests to ensure the cohort random split has not introduced a significant difference between the train, validation and test splits using ANOVA test for age and Chi-Square test for categorical variables. The p-values are reported in the final column.

Variable	Train	Validation	Test	All	P-value
N	30065	7502	9394	46961	
Age (mean ± std)	53.7 ± 13.0	53.6 ± 13.3	53.8 ± 13.2	53.7 ± 13.1	0.4975
Female	17369 (57.8%)	4289 (57.2%)	5417 (57.7%)	27075 (57.7%)	0.6421
Race: White	13607 (45.3%)	3453 (46.0%)	4299 (45.8%)	21359 (45.5%)	0.4054
Race: Black or African American	7713 (25.7%)	1873 (25.0%)	2503 (26.6%)	12089 (25.7%)	0.0391
Race: Asian	569 (1.9%)	146 (1.9%)	169 (1.8%)	884 (1.9%)	0.7651
Race: Other	8176 (27.2%)	2030 (27.1%)	2423 (25.8%)	12629 (26.9%)	0.0263
Progressed to CKD	2803 (9.3%)	666 (8.9%)	911 (9.7%)	4380 (9.3%)	0.1903

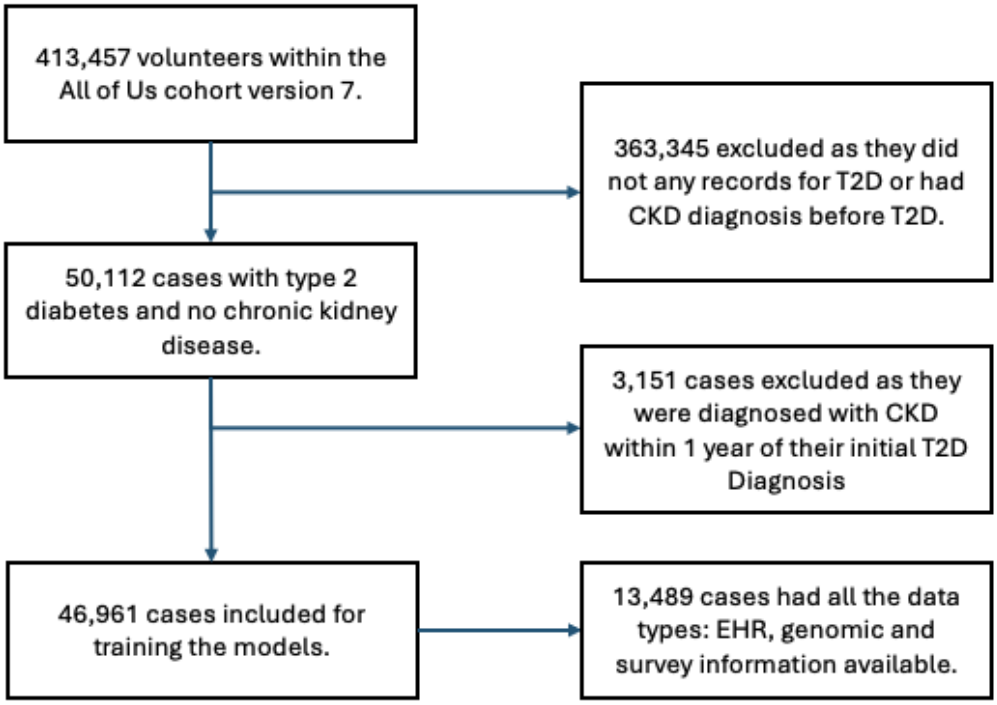

Figure 3.2. Cohort flowchart. From the original 413,457 volunteers in All of Us, 50,112 had type 2 diabetes and no chronic kidney disease.

3.5.2 Polygenic Risk Scores

Whole genome sequencing data was available for 32,808 volunteers within the study cohort. The table below summarizes the PRS scores for the volunteers grouped by their CKD status. Both PRS scores were significantly associated with the outcome. **Table 3.3** summarizes the polygenic risk scores and their association with progression to CKD.

Table 3.3. The polygenic risk scores (PRS) are summarized. A t-test was used to check for statistically significant difference between the cohort that progressed to CKD and the cohort with no CKD diagnosis.

PRS	Number of SNPs	Mean for patients with CKD (n = 2,837)	Mean for patients without CKD (n = 29,935)	P-value (t-test)
1	41,419	-0.307	-0.332	<1e-6
2	158	0.068	0.064	0.02

3.5.3 Survey Information

The survey information was available for 37,545 participants. If some survey questions or surveys were missing from any participant, they were replaced with a placeholder value for missing values.

3.5.4 Transformer Model Pretraining

The masked concept modeling approach described above was used to pretrain the base transformer model. The RoBERTa based model was altered to accept a vocabulary size of 20695. The masked

concept modeling approach was performed for 20 epochs and the final validation loss declined from 9.94 to 2.80. This step of the training took 12 hours to conclude using a NVIDIA T4 GPU.

3.5.5 Unimodal Model Performance

Among the unimodal models, the transformer model using the EHR data had the highest performance achieving an AUROC of 0.73 (95% CI: 0.71 – 0.75).

The survey-based model achieved an AUROC of 0.59 (0.56 – 0.60) on the test set. It achieved an average AUROC of 0.61 on cross-validation. After adjustment for age and sex, the most important predictors were variables related to alcohol consumption levels and tobacco use. Among the SDOH variables, the lower score for social support scale was associated with progression to CKD. (**Figures 3.3 and 3.4**)

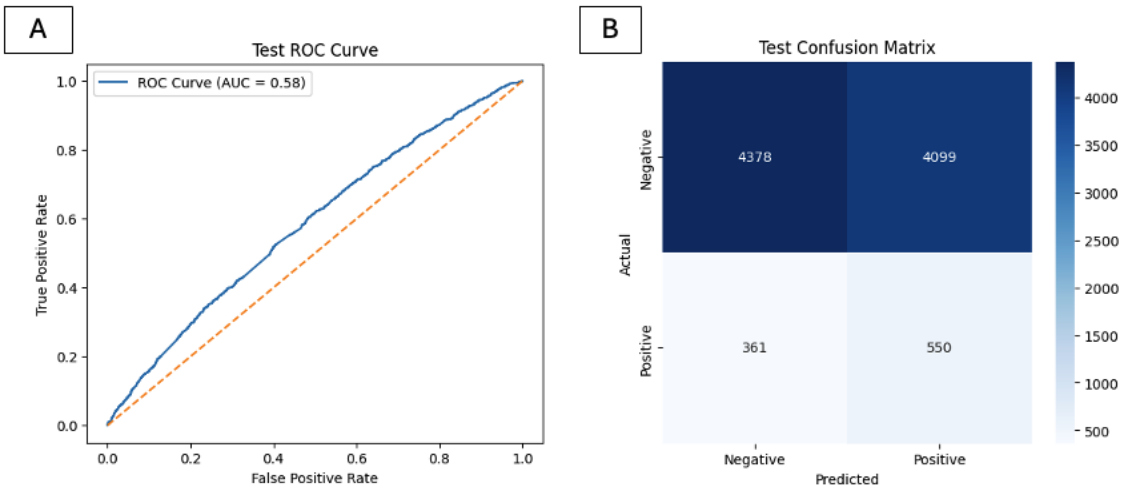

Figure 3.3. Receiver operator characteristic (ROC) curve and the confusion matrix for the survey-based model over the test set.

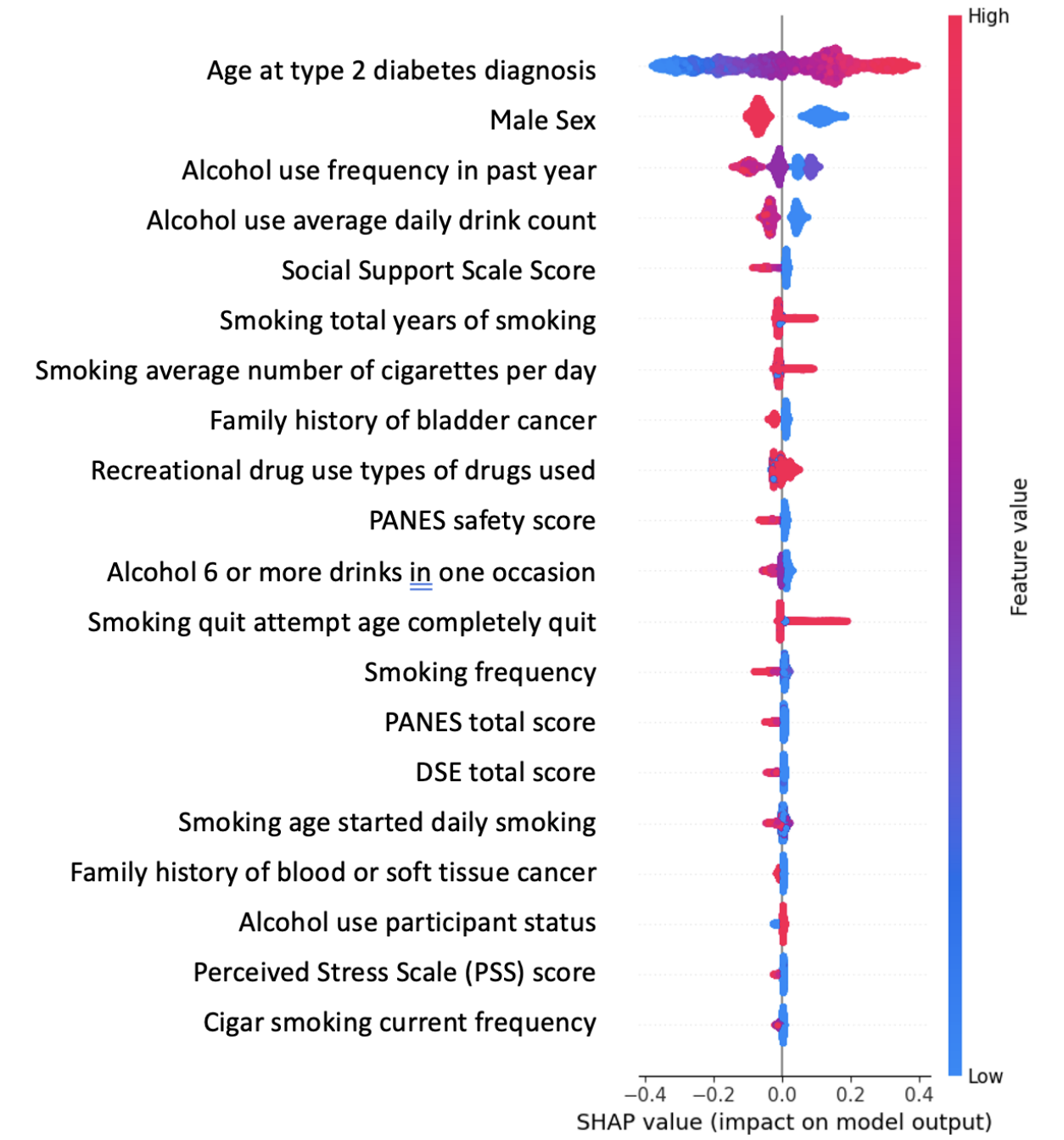

Figure 3.4. SHAP plot for feature importance in the survey model, adjusted for age at the time of diagnosis and sex.

The genomic based XGBoost model reached an AUROC of 0.53 (0.50 – 0.55) on the test set. It achieved a AUROC of 0.57 on cross-validation. Using shapely values, the PRS scores were the most important features in prediction of progression to CKD, with the larger PRS contributing the most to the prediction. 19 other SNPs had non-zero contribution to the model predictions. Multiple of these SNPs belonged to the chromosome 12, focused in the various regions associated with the Centrosomal protein of 290 kDa (CEP290) which has been associated with medullary cystic kidney disease. The other gene implicated in our analysis based on the explainability analysis was the nephrocystin 1 (NPHP1) gene in chromosome 2. Multiple SNPs within the gene location were among the selected features in the genomic model. (**Figures 3.5 and 3.6**)

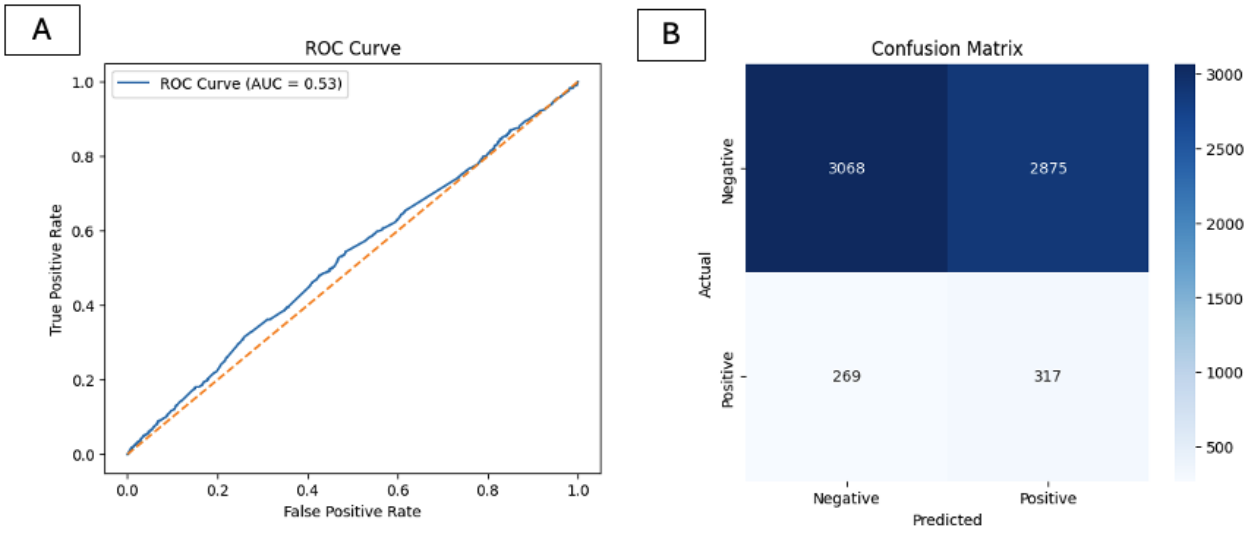

Figure 3.5 Receiver operator characteristic (ROC) curve and the confusion matrix for the genomic model over the test set. The genomic model contained both the PRSs and the list of pathogenic and likely pathogenic SNPs from ClinVar.

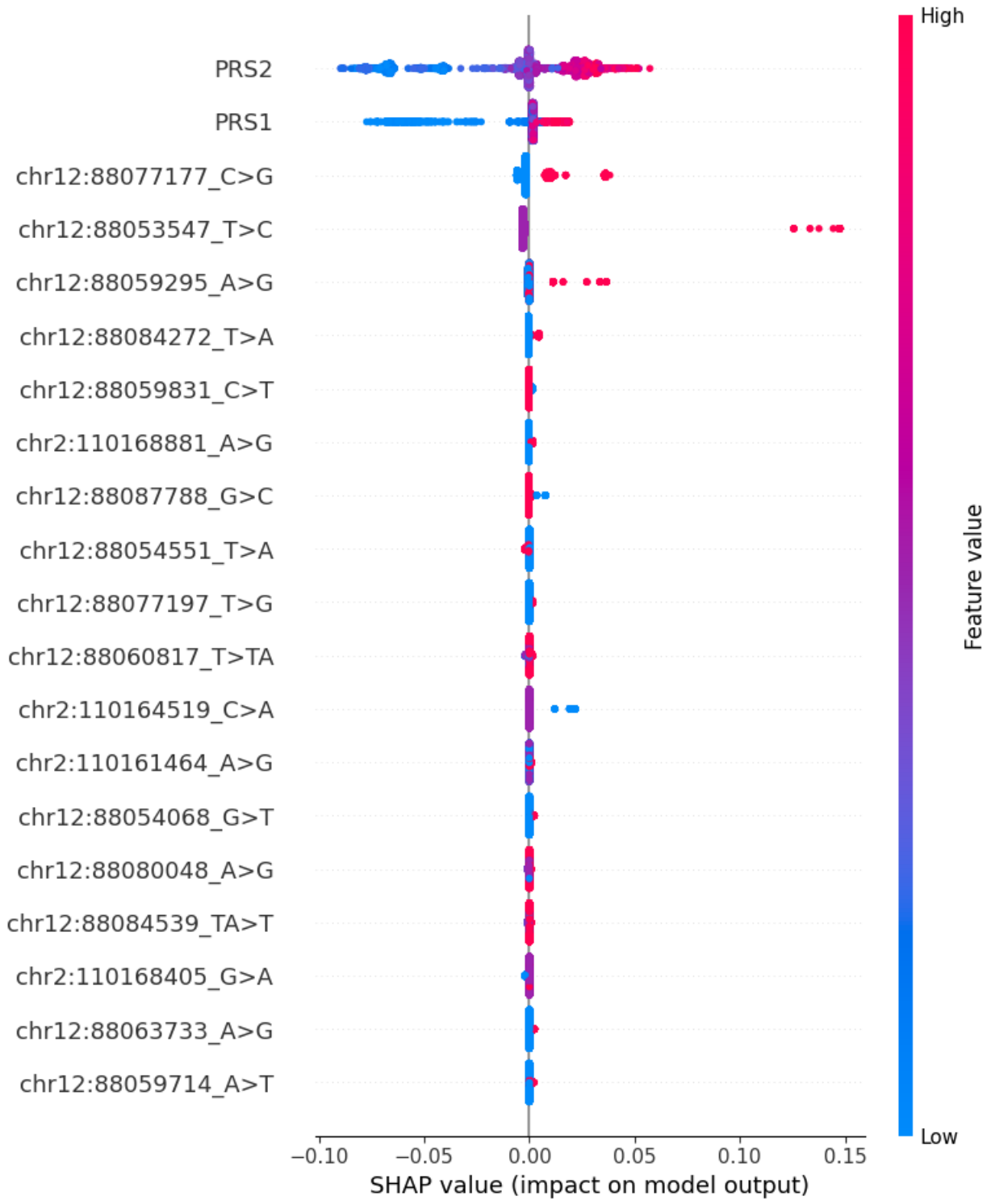

Figure 3.6. SHAP plot for feature importance in the genomic. Only 22 variables had none-zero contribution to the regularized model. The two PRSs had the most impact on model prediction.

3.5.6 Multimodal Model Performance

While based on validation set performance the best performing model was the multimodal model using the early fusion strategy with a validation AUROC of 0.74 (0.72 – 0.76), this was not significantly higher than other fusion strategies or the unimodal EHR transformer. This model achieved an AUROC of 0.72 (0.70 – 0.74) on the hold-out test. The late fusion model followed achieving an AUROC of 0.73 (0.71 – 0.75) on the validation set and a similar AUROC of 0.73 (0.71 – 0.75) on the test set. The full details of all model performance metrics can be viewed in **Table 3.4.**

Table 3.4. Summary performance metrics for the unimodal and multimodal models across the train, validation, and testing splits.

Model Type	Test AUROC	Validation AUROC	Train AUROC	Test Sensitivity	Test Specificity	Test PPV	Test NPV
EHR Transformer	0.73 (0.71 – 0.75)	0.73 (0.71 – 0.75)	0.80 (0.79 – 0.81)	66%	68%	18%	95%
Genomics (XGBoost)	0.53 (0.50 – 0.55)	0.56 (0.53 – 0.59)	0.55 (0.54 – 0.57)	54%	52%	10%	92%
Survey (XGBoost)	0.59 (0.56 – 0.60)	0.61 (0.58 – 0.63)	0.62 (0.61 – 0.63)	60%	52%	12%	92%
	Multimodal Models						
Early Fusion	0.72 (0.70 – 0.74)	0.0.74 (0.72–- 0.76)	0. 79 (0.77 – 0.80)	49%	83%	24%	94%
Intermediate Fusion	0.72 (0.71 – 0.74)	0.72 (0.70 – 0.74)	0.81 (0.80 – 0.82)	61%	74%	20%	95%
Late Fusion	0.73 (0.71 – 0.75)	0.73 (0.71 – 0.75)	0.80 (0.79 – 0.80)	67%	65%	18%	95%

3.5.7 Fairness Analysis

For fairness analysis, we tested the performance of the early fusion model as it was selected as the best performing model using the validation set. The performance metrics can be seen in **Table 3.5.**

Table 3.5 Fairness analysis showing model performance across subpopulations in the study.

Split	Count	AUROC	AUPRC	Sensitivity	Specificity	Positive Predictive Value	Negative Predictive Value
Race: White	21305	0.7 (0.68–0.73)	22%	75%	55%	16%	95%
Race: Black or African American	12035	0.72 (0.70–0.76)	29%	70%	64%	18%	95%
Race: Other	12345	0.73 (0.69– 0. 76)	27%	54%	82%	22%	95%
Race: Asian	881	0.72 (0.55–0.88)	19%	82%	61%	13%	98%
Race: Middle Eastern or North African	219	0.69 (0.19–1.00)	21%	50%	94%	33%	97%
Sex: Female	26940	0.72 (0.70–0.75)	23%	65%	69%	16%	96%
Sex: Male	18813	0.69 (0.66–0.71)	24%	59%	68%	19%	93%
Age: 18-29	2036	0.7 (0.53–0.88)	20%	50%	87%	10%	98%
Age: 30-39	4685	0.75 (0.68–0.82)	23%	72%	69%	12%	98%
Age: 40-49	9327	0.7 (0.66–0.75)	24%	46%	86%	21%	95%
Age: 50-59	14407	0.72 (0.69–0.76)	25%	80%	54%	16%	96%
Age: 60-69	11495	0.7 (0.66–0.73)	25%	57%	75%	24%	93%
Age: 70-79	4201	0.6 (0.54–0.66)	19%	32%	85%	25%	89%
Age: 80+	634	0.63 (0.49–0.80)	41%	44%	87%	44%	87%

3.6 Discussion

In this aim, titled “Evaluation and comparison of early, intermediate, and late fusion techniques for combining exposures, clinical and genomics data for disease risk prediction task using All of Us: Risk of CKD in patients with type 2 diabetes”, we applied multimodal modeling approaches to EHR, genomic and survey data to develop multimodal models for early prediction of progression to CKD in patients with T2D. We tried to answer part of our over-arching question on evaluation of multi-modal data fusion approaches for predictive clinical models by combining EHR, survey and genomic data. This is the first model that integrates all these data types for prediction of CKD in T2D. We demonstrated how the addition of the survey and genomic data can improve the performance of the model. However, this improvement was minimal due to the limited ability of data extracted from these modalities to predict progression to CKD in patients with T2D.

Compared to previous efforts, our model has multiple distinct features. First, to the best of my knowledge, this is the first model combining EHR, genomics and survey information (SDOH, lifestyle and family history data) to predict progression to CKD. In addition, we aimed for early prediction of patient progress to CKD within the next 6 years after their original diagnosis of T2D (excluding those diagnosed within the first year after T2D diagnosis). Most other studies focused on prediction of CKD in patients in any stage of the T2D. Our task is inherently more challenging as patients with longer history of T2D are more likely to demonstrate signs of kidney damage and other related adverse outcome like optical retinopathy (26), and will have gone through more work up as part of the T2D routine clinical care. Therefore, our model has the benefit of identifying those at a high risk earlier which enables the targeted use of preventive measures for these individuals before signs of kidney disease appear (27,28). Compared to previous efforts, our model

has multiple distinct features. First, to the best of my knowledge, this is the first model combining EHR, genomics and survey information (SDOH, lifestyle and family history data) to predict progression to CKD. In addition, we aimed for early prediction of patient progress to CKD within the next 6 years after their original diagnosis of T2D. Most other studies focused on prediction of CKD in patients in any stages of the disease. Our task is inherently more challenging as patients with longer history of T2D are more likely to demonstrate signs of kidney damage and other related adverse outcome like optical retinopathy (26) and will have gone through more work up as part of the T2D routine clinical care. However, our model has the benefit of identifying those at a high risk earlier which enables the targeted use of preventive measures for these individuals before signs of kidney disease appear (27,28).

As we hypothesized, the use of multimodal patient data, combined with longitudinal EHR data, improved the performance of our models compared to our unimodal models. However, the magnitude of improvement was small. Multiple factors could have contributed to this observation. Despite the statistically significant relationship between the calculated polygenic risk scores and progression to chronic kidney disease, the magnitude of that relationship was small. Despite our efforts to only include PGSs that had been validated on some diverse populations, the PGSs used in this study were developed in populations with mostly European ancestry. Using ancestry specific PGSs, if available, could mitigate this problem. Additionally, although genetics is important in chronic kidney disease, studies have found that the associations are usually weak (35). Like the genomic model, the model developed using the survey information did not have high discrimination power between cases and controls, achieving an AUROC of only 0.57 over the test set. Additionally, the predictive value of family history, lifestyle and SDOH conditions and genetic make-up may already be reflected by other observations from the EHR (e.g. chronic obstructive

pulmonary disease in people who smoke). As we hypothesized, the use of multimodal patient data combined with longitudinal EHR data improved the performance of our models compared to our unimodal models. However, the magnitude of improvement was small. Multiple factors could have contributed to this observation. Despite the statistically significant relationship between the calculated polygenic risk scores and progression to chronic kidney disease, the magnitude of that relationship was small. Despite our efforts to only include PGSs that were at least validated on diverse populations, the PGSs used in this study were developed in populations with mostly European ancestry. Using ancestry specific PGSs if available could mitigate this problem. Additionally, although genetics is important in chronic kidney disease, studies have found that the associations are usually weak (35). Like the genomic model, the model developed using the survey information did not have high a discrimination power between cases and controls achieving an AUROC of 0.57 over the test set. Adding this to the fact that some of the implications of both the family history, lifestyle and SDOH conditions of the patients, and their genetic make-up may already appear in the EHR (e.g. chronic obstructive pulmonary disease in people who smoke).

We explored three main fusion strategies for combining genomic, EHR and survey data. Based on validation set performance, the early fusion strategy slightly outperformed the other fusion strategies and the unimodal transformer model. However, this difference was not statistically significant and was not replicated in the hold out test set. Additionally, the model using the early fusion strategy had the least level of overfitting to the training set. The very small improvement in model performance when additional modalities were added makes it difficult to assess the utility of these three fusion strategies in the case of prediction of progression to CKD.

Despite the negative findings, our study shed light on some aspects of the multimodal data fusion specifically when using transformer models. The early fusion network had the least number of

parameters. In addition, by directly adding the positive SNPs, PRS scores and survey questions and filtering out the negative ones, the survey information could be passed into the model without losing potential signal. This is a limitation of early fusion when the data types are less compatible (e.g. imaging and longitudinal EHR data). The lower extent of overfitting observed in early fusion could potentially be attributed to the lower number of trainable parameters.

A benefit of Late fusion that was observed in this study was the ability to ensemble different types of models (e.g. tree based XGBoost and transformer). This enables using the best architecture for each data type. Since these models train using different approaches, this is not as easy to implement in intermediate fusion and not possible in early fusion.

Our fairness analysis showed that the performance of our model varied in different subpopulations. Notably, the model had a higher performance predicting progression to CKD in Female patients and the 30-39 age group. Additionally, the performance was much lower for the patients older than 70 years old. This could highlight problems with follow up, competing outcomes or inherent data distribution issues in those categories. Additionally, the lower number of patients in this age group may have contributed to the lower performance.

3.6.1 Limitations

We identified multiple limitations for our study some of which could be avenues for future work. While there exists an increasing trend to analyze longitudinal EHR data using transformer models (37), there are multiple design choices that need to be made that have unclear influence on the final performance of the model. Notably, while we decided to include all concept ids within the EHR data regardless of their frequency as distinct tokens, it is possible to aggregate concept ids with

lower frequencies to parent concepts. Additionally, increasing the maximum model input length can potentially improve performance at the cost of the complexity of the model.

For genomic information, we used external knowledge injection to extract a list of relevant pathogenic variants from ClinVar and two polygenic risk scores. Although we demonstrated that the polygenic risk scores were significantly associated with the outcome, the magnitude of this association was small. A limitation of this approach is that those pathogenic variants and polygenic risk scores were developed on populations that had different ancestry distributions compared to the All of Us dataset (36,45). The two risk scores were developed on data from populations with mostly European ancestries (although they were validated on more diverse populations.). For the genomic information, we used external knowledge injection to extract a list of relevant pathogenic variants from ClinVar and two polygenic risk scores. Although we demonstrated that the polygenic risk scores were significantly associated with the outcome, the magnitude of this association was small. A limitation of this approach is that those pathogenic variants and polygenic risk scores were developed on populations that had different ancestry distributions compared to the All of Us dataset (36,45). The two risk scores were developed on datasets of mostly European ancestries (although they were validated on more diverse populations.).

3.7 Conclusion

In this aim, to address our overarching question of evaluation of multimodal data fusion approaches in predictive medical models, we proposed an approach to integrated genomic, survey information and longitudinal EHR data into deep learning models. Our deep learning model enables early identification of individuals with T2D that are at a high risk for progression to CKD. While we demonstrated that multiple genomic locations and polygenic risk scores are associated with progression to CKD, the addition of these modalities into longitudinal EHR data did not

significantly improve model predictions. To answer our overarching question, we explored the three main data fusion approaches, early, intermediate, and late fusion. Although our findings were inconclusive about the best performing fusion approach in this scenario, we highlighted some of the benefits of using each of the fusion strategies when combining genomic, EHR and survey data. Given the nature of these data types, they could be passed to the early fusion model without much alteration that may result in loss of valuable information. The early fusion strategy had the least number of parameters and suffered from the least amount of overfitting. On the other hand, the late fusion strategy enabled modality specific modeling strategies that could not be integrated and trained together easily. In our late fusion model, we easily combined two XGBoost models with a transformer. This could be valuable when one specific modeling approach is superior to other approaches. Over the next two chapters, I will explore the problem of fusion even further, introducing new data modalities (Imaging, structured EHR data), modeling approaches (2D and 3D convolutional neural networks) and outcome types. Lessons learned from this aim about fusion can be seen in **Figure 3.7**.

Early Fusion	Intermediate Fusion	Late Fusion
• Good choice if features can be represented with no simplification/dimensionality reduction. • Can handle missing data if appropriate architecture is used.	• May overfit to training data due to large number of trainable parameters.	• Can combine diverse networks like XGBoost and transformer.

Figure 3.7 Lessons learned about fusion in Aim 1.

4. Aim 2: Development and assessment of the incremental value of combining a deep convolutional neural network (CNN) feature extractor on imaging data and clinical data on a binary prediction task: Predict post-surgical margin status in soft tissue sarcoma.

In the previous chapter, we explored methods to use early, intermediate, and late fusion strategies to combine EHR, genomic and survey data using transformer models. As described in the background and significance (Chapter 2), medical data includes other important modalities including imaging that pose challenges during multimodal integration. To further explore the implications of multimodal data fusion in predictive modeling in medicine, in this chapter, I will extend the concepts described earlier in Chapter 2 and Chapter 3 by including additional medical data domains and deep learning network architectures. While the previous chapter was mainly focused on fusion of EHR, genomics and survey data and used a combination of fully connected neural networks and state of the art transformer models to analyze those data; in this aim, I will explore the added value of combining cross-sectional multiparametric imaging data (magnetic resonance imaging) with a structured set of clinical and pathology features extracted from the electronic health records. I will also look at the impact of early, intermediate, and late fusion strategies on the multi-modal model performance in combing these data types, specifically in the context of using convolutional neural networks is analyzing the imaging data.

4.1 Background:

Soft tissue sarcoma (STS) is a rare type of cancer often involving the extremities or trunk. They can originate from tissues like muscle, connective tissue, and adipose tissue and are highly diverse (52). Their rarity and diversity have made developing effective treatment strategies more challenging. Hence, using personalized medicine strategies to optimize care in patients with STS, based on their personal medical history and their cancer subtype could yield promising results.

Typically, STSs are treated by neoadjuvant radiotherapy with or without chemotherapy or more recently immunotherapy followed by surgical excision (53–55). One of the most important goals of surgery is preserving as much function as possible (53). Previous studies have cited multiple prognostic factors in STS including tumor size, grading, and margin status (56–58). Achieving post-surgical margin control is one of the most important predictors of recurrence and disease-free survival in patients with STS (57–62). In fact, in some studies, margin was the sole important predictor of local recurrence of STS surpassing other variables like tumor viability, neoadjuvant treatment, grading, and size (63). There exists a lack of consensus on the depth of clear margin required to avoid local recurrence, but most studies have shown that anything between no viable tumor on margins to 1-2mm of clear margins show the most significant benefit (61,64). Hence, surgeons need to strike a fine balance between removing enough tissue to achieve clear margins and avoiding unnecessary loss of function due to radical excision. Despite best efforts, about 15-25% of STS surgeries result in positive or close margins (65). Strategies to identify patients at a higher risk for positive margins can help the care team with treatment planning (64).

Previous studies have introduced radiologic signs that are associated with tumor marginal infiltration, but there is no established risk model. Signs such as infiltrative growth pattern (66), peritumoral contrast enhancement and existence of tail sign (64,67) may be indicative of an increased risk for positive tissue margins (64,68). Machine learning approaches have been previously used successfully to predict post-surgical margin status in other cancers. Radiomics analysis has also been used in STS to predict margin status on Dixon MRI (69).

4.2 Significance

Given my overarching question of evaluating multimodal data fusion approaches in predictive medical models, and the gap in literature regarding the utility of different fusion approaches in combining imaging data with clinical data, this aim will look into early, intermediate and late fusion in the context of integrating longitudinal cross sectional imaging data with clinical. The outcome of interest in this question is a binary outcome compared to the next chapter which looks at combining 2D imaging and clinical data to solve a continuous regression problem.

In this study, we used multimodal deep learning approaches using multiple fusion strategies to combine longitudinal MRIs with relevant clinical and pathology information to develop a deep learning model for prediction of post-surgical margin involvement in patients with STS. By incorporating information like the histopathologic type and grade of the tumor, we expect that our model will output histology specific risk scores. Additionally, we developed another explainable model based on radiologist evaluation of the longitudinal MRIs as baseline for comparison. By comparing the performance of the multimodal models across the three fusion strategies in prediction of margin status in STS, we will shed light on the implications of fusion on binary prediction tasks when combining imaging and clinical data. In the next section, I will go through the methods used to prepare the dataset and develop these deep learning models.

4.3 Materials and Methods:

This study was conducted in accordance with the Checklist for Artificial Intelligence in Medical Imaging (CLAIM) guidelines (70). Institutional review board approval was obtained (IRB#STUDY00011377), with a waiver of informed consent granted for the retrospective analysis of anonymized patient data.

4.3.1 Data Sources and Study Population:

We identified eligible patients from our institution's patient data repository, selecting those diagnosed with primary STS who underwent neoadjuvant radiotherapy with or without chemotherapy between 2008 and 2021. Inclusion criteria required the availability of pre-treatment MRI scans and post-excision pathology reports. A detailed cohort flowchart is provided in **Figure 4.1**. Cohort characteristics can be viewed in **Table 4.1**.

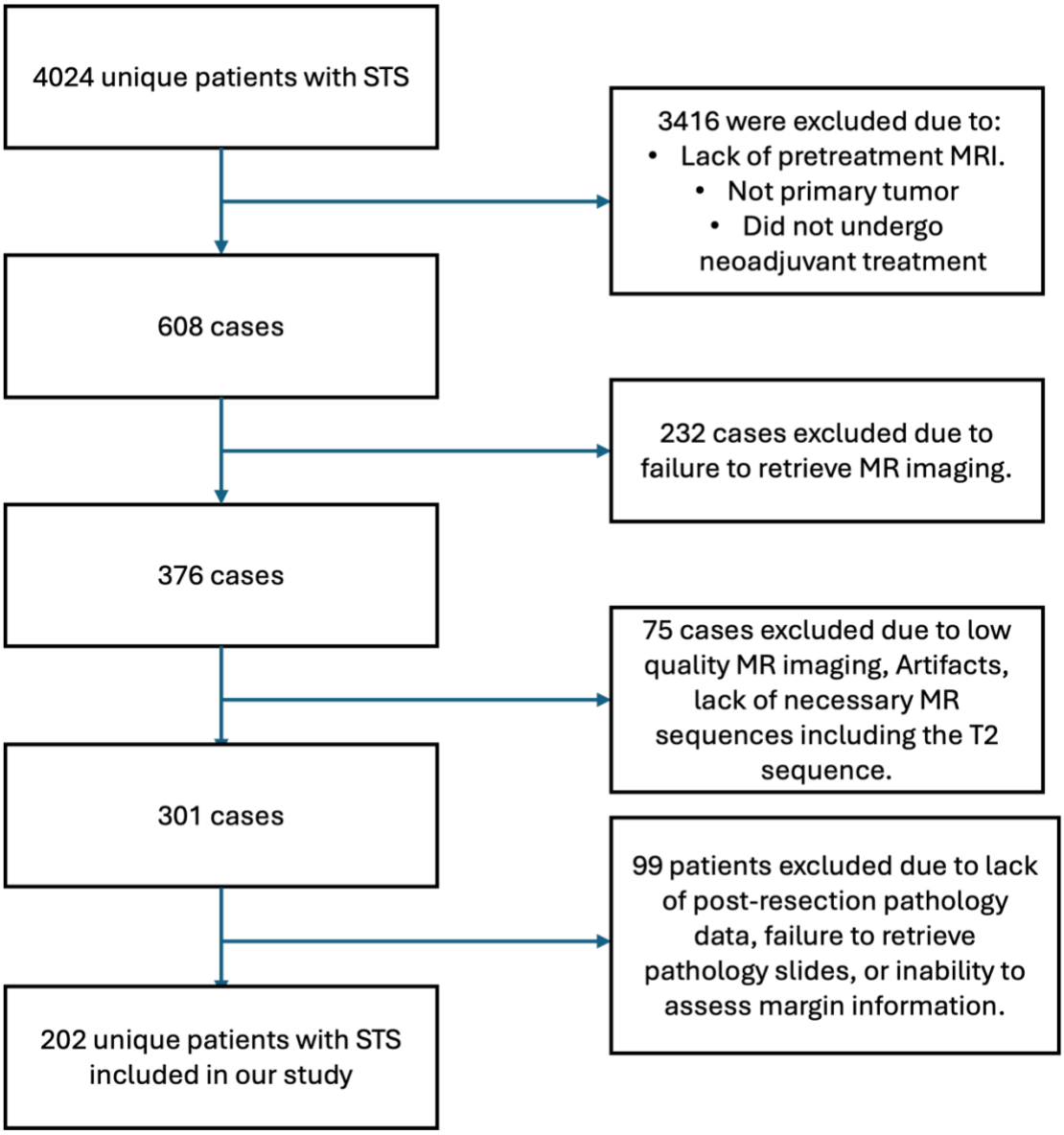

Figure 4.1: Cohort Flowchart. Our initial query revealed 4024 unique patients treated at our center of which after application of all inclusion and exclusion criteria and data retrieval steps, 202 qualified for the final analysis.

Table 4.1 Details of cohort characteristics in the training and testing sets.

	All (%)	Involved Margin (%)	Clear Margins (%)	p-Value
Age	53.4 ± 15.6	54.7 ± 17.1	53.0 ± 15.1	0.543
Female	74 (36.6%)	21 (40.4%)	53 (35.3%)	0.525
Male	128 (63.4%)	31 (59.6%)	97 (64.7%)	0.525
Grading				
High Grade	62 (30.7%)	13 (25.0%)	49 (32.7%)	0.288
Intermediate Grade	37 (18.3%)	10 (19.2%)	27 (18.0%)	0.847
Low Grade	20 (9.9%)	3 (5.8%)	17 (11.3%)	0.185
Grade Not Defined	74 (36.6%)	24 (46.2%)	50 (33.3%)	0.112
Location				
Trunk	31 (15.3%)	12 (23.1%)	19 (12.7%)	0.113
Head & Neck	1 (0.5%)	1 (1.9%)	0 (0.0%)	0.322
Lower Extremity	134 (66.3%)	27 (51.9%)	107 (71.3%)	0.016
Upper Extremity	34 (16.8%)	11 (21.2%)	23 (15.3%)	0.368
Histopathology				
Tumors of uncertain differentiation	85 (42.1%)	19 (36.5%)	66 (44.0%)	0.346
Fibroblastic/Myofibroblastic tumors	25 (12.4%)	6 (11.5%)	19 (12.7%)	0.83
Adipocytic tumors	26 (12.9%)	5 (9.6%)	21 (14.0%)	0.384
Smooth muscle tumors	14 (6.9%)	3 (5.8%)	11 (7.3%)	0.689
Other Subtypes	52 (25.7%)	19 (36.5%)	33 (22.0%)	0.058
Post-surgical Margin	25.70%			

4.3.2 MRI Acquisition:

The MRI protocol included multiple axial sequences: T2-weighted spin-echo fat-saturated (T2), T1-weighted spin-echo fat-saturated images acquired before (T1BC) and after (T1AC) contrast administration, and non-fat-saturated T1-weighted spin-echo images (T1). For the T2 sequence, echo time (TE) and repetition time (TR) ranged from 60 to 120 ms and 2000 to 8000 ms, respectively. For T1AC, TE ranged from 2 to 25 ms and TR from 350 to 1000 ms. T1BC parameters included a TE range of 6 to 20 ms and a TR range of 350 to 800 ms. The T1 sequence exhibited

TE values between 2 and 25 ms and TR values between 350 and 1200 ms. Slice thickness across all sequences varied from 1 to 8 mm.

4.3.3 Input Data:

Tumor volumes were manually delineated on all axial slices by a fellowship-trained musculoskeletal radiologist with seven years of post-fellowship experience (M.C.), utilizing both pretreatment and posttreatment MRI scans (where applicable) across all relevant imaging sequences. Manual segmentations were performed using MiM Software (version 7.1.2; MIM Software Inc., Cleveland, OH, USA). A second musculoskeletal radiologist (C.P.), also with seven years of post-fellowship experience, reviewed all segmentations for accuracy and consistency. Any discrepancies between the two readers were resolved by consensus. To assess interobserver reliability, the second radiologist independently segmented a randomly selected subset of 50 MRI studies. Given potential variations in patient positioning and anatomical alignment across different sequences, tumor segmentation was performed independently for each MRI sequence. All tumor regions identified on axial slices were included in the final analysis. After annotation, a bounding box was drawn around each tumor volume with 1 cm of additional space around the tumor. The images were cropped using this bounding box.

Image preprocessing steps included resampling the cross-sectional images to a voxel size of 1x1x1 (mm), voxel intensity normalization to values between 0 and 1, and resizing the cropped images to 4*16*128*128. The pre-treatment and post-treatment images were concatenated on the channel level to generate the final input image with dimensions of 8*16*128*128. Missing sequences were replaced with an image of the same size will voxel values equal to zero.

Clinical variables included in this study were manually extracted from the electronic health records. Variable selection was guided by a comprehensive review of the relevant literature and expert consensus. The extracted variables comprised tumor grade, histopathologic subtype, patient age at diagnosis, sex, and tumor anatomical location (**Table 4.2**).

4.3.4 Radiologist Evaluations and Semantics Features:

A fellowship-trained musculoskeletal radiologist (C.P.), with seven years of experience, and another radiologist (S.H.), with five years of experience, independently reviewed all MRI scans to extract a predefined set of radiologic features (semantic features) important in soft tissue sarcoma assessment including factors associated with infiltrative disease including pre-tumoral contrast enhancement and tail sign. These features were selected based on previous studies (65) and expert consensus and are detailed in **Table 4.2**.

Table 4.2: List of features extracted by the radiologists for evaluation of post-surgical margin. Two independent readers evaluated each imaging, and the agreement is recorded here.

Semantic Feature	Values	Inter-reader agreement (weighted Kappa)
Location	Subcutaneous, Deep Fascia, Muscle, Intermuscular	0.92
Confinement	Confined, Not Confined	0.86
T1 margin	> 90%, 90-50%, < 50%	0.91
T2 margin	> 90%, 90-50%, < 50%	0.89
Contrast enhanced margin	> 90%, 90-50%, < 50%	0.90
T1 heterogenous signal intensity	< 50%, ≥50%	0.89
T2 heterogenous signal intensity	< 50%, ≥50%	0.91
Contrast enhanced heterogenous signal intensity	< 50%, ≥50%	0.89
Intratumoral enhancement	Absent, Present	0.97
Necrosis	No Necrosis, ≤50%, >50%	0.85
Intralesional fat	Yes, No	0.95
Hemorrhage	Yes, No	0.94

Peritumoral enhancement	No, Limited, Extensive	0.93
Tail sign	Yes, No	0.91
Vessel encasement	Yes, No	0.93
Bone involvement	No, Encasement, Invasion	0.98
Size	Centimeter of largest diameter	0.72

4.3.5 Outcome definition

The primary outcome of our study was based on whether the post-surgical tissue margins contained viable tumor tissue or were clear. A positive outcome was defined as when there was involvement of post-surgical margin.

4.3.6 Longitudinal vs. Single Time Point

While all subjects in our dataset had pretreatment MR imaging, only a subset had longitudinal MRIs available. After splitting the data into training and test sets, we designated the subset of the test set with longitudinal imaging as our longitudinal test set. This approach allowed us to maximize the use of cases with only pretreatment images while preserving the ability to evaluate our model's performance on longitudinal cases.

4.3.7 Modeling Approach and Fusion Strategies:

For the imaging part of the model, we used a pretrained 3D ResNet10 architecture provided by MONAI altered to accept 3D images with 8 channels. We also experimented with 3D ResNet50 and 3D ResNet101 models as hyperparameters but eventually decided to use the ResNet10 model

to balance model complexity and its potential for overfitting. For the clinical portion of the data, we used a shallow neural network with 2 fully connected layers.

The multimodal nature of the input data necessitates implementing strategies to combine the clinical and imaging information together. These strategies, often referred to as fusion strategies, involve extracting information from each modality and bringing them into a compatible form. Subsequently, the extracted information is combined, and a final prediction is made. For our multimodal deep learning model, we combined the two networks using three main fusion strategies: early, intermediate and late.

Early fusion strategy involves bringing the multiple data types into a compatible format and combining them before passing them into the neural network. For ealy fusion, we used the pretrained 3D ResNet10 model as a feature generator with fixed weights and used the final embedding layer as the input to the multimodal model. We concatenated these embeddings with the clinical variables and passed them through a fully connected neural network with 3 layers to make the final prediction (**Figure 4.2**).

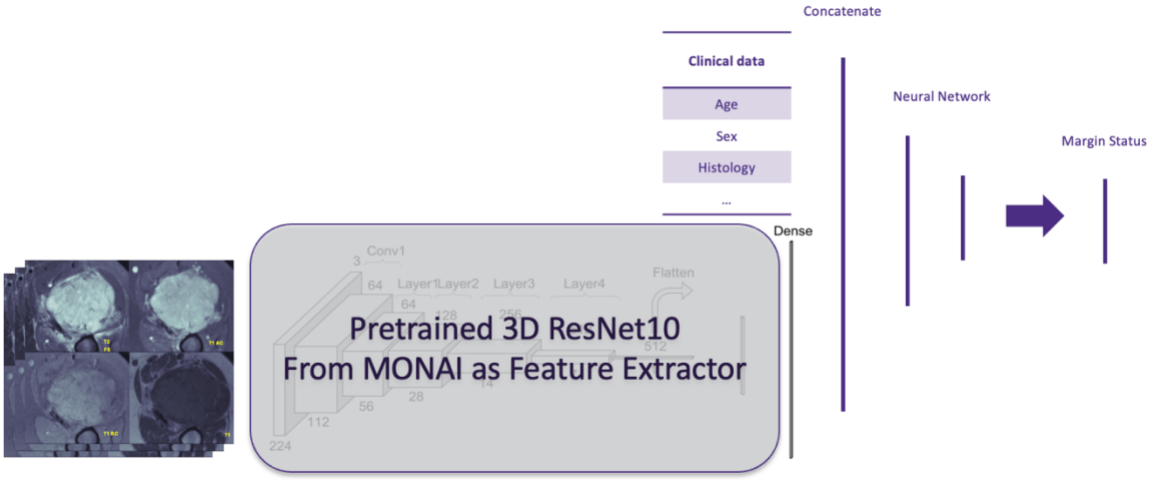

Figure 4.2 Early fusion network architecture. In this network, a pretrained 3D ResNet10 model is used to generate features from the imaging data that are then concatenated with the clinical variables to make the final prediction.

For the intermediate fusion strategy, we combined the imaging and clinical model by concatenating their final embedding layer and adding a fully connected neural network layer at the end to make the final prediction. In this variation, both the 3D ResNet10 model and the clinical model are fined tuned during the model training stage (**Figure 4.3**).

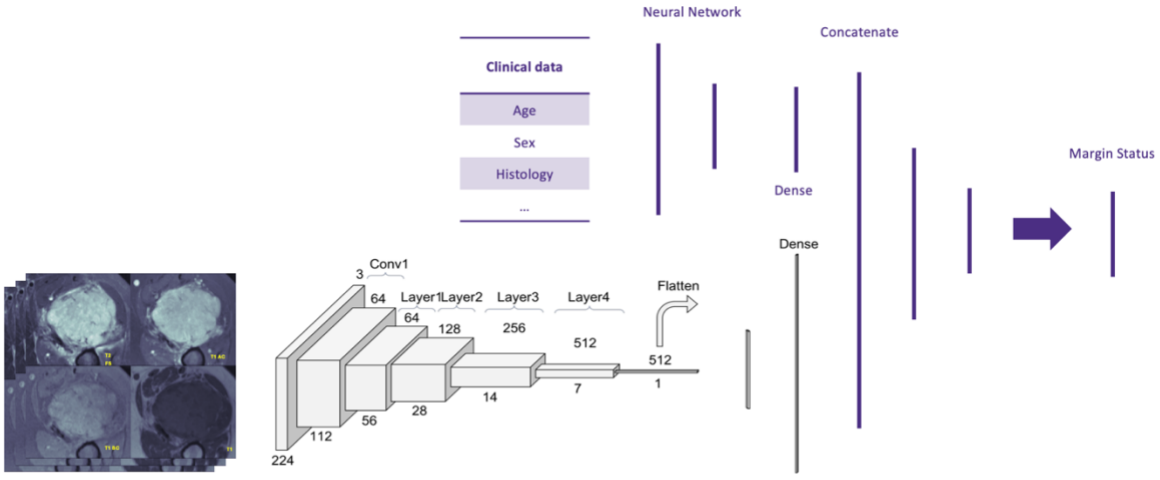

Figure 4.3 Intermediate fusion network architecture.

For the late fusion strategy, we fully trained an imaging only model and a clinical only model. Subsequently, we used fully connected neural network that took the predictions of each of these unimodal models and made a final prediction based on those (**Figure 4.4**).

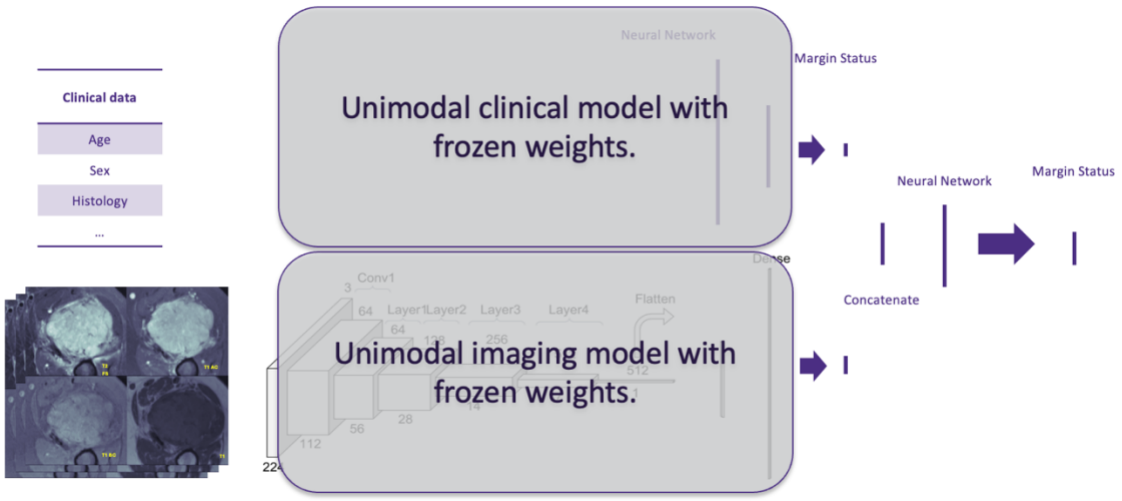

Figure 4.4 Late fusion network architecture.

We split the data into train and test sets with a 0.8/0.2 ratio on a patient level. 3-fold cross-validation over the test set was used to identify the best set of hyper parameters. The hyper-parameters that we explored included the imaging network architecture (3D ResNet10, 3D ResNet50 and 3D ResNet101), the fusion timing, learning rate and weight decay. Grid search was performed over the hyperparameter space to identify the optimum values.

4.3.8 Performance metrics:

We evaluated each model's performance by calculating its average area under the receiver operator curve (AUROC) during classification. After identifying the best hyperparameters for each model, the final model was tested on the independent test set. The confidence intervals for AUROC was calculated in the train and test sets using the DeLong method (71). In addition, we reported sensitivity and specificity over the test set. Finally, we tested our models on the portion of the test set that included both pre and post-treatment longitudinal images.

4.3.9 Explainability

For the clinical and pathology data, we used Shapley explanations to generate feature importance plots. For the imaging section, we used occlusion maps and integrated gradients.

4.3.10 Statistical Analysis

Model performance was evaluated using the area under the receiver operating characteristic curve (AUROC). The DeLong method was used to compute confidence intervals for performance metrics in the testing set (71). The Youden Index was used to identify the best threshold for prediction over the training set. During cross-validation, the average AUROC was used to determine the optimal set of hyperparameters. Interobserver agreement for ROI segmentation masks was assessed using the Dice similarity coefficient. A p-value of <0.05 was considered indicative of statistical significance.

4.4 Results:

4.4.1 Cohort Characteristics:

The final cohort consisted of 202 patients with soft tissue sarcoma. After random splitting, 159 patients were assigned to the training set, while the testing set comprised 43 patients. The mean age at diagnosis was 54 years (SD 15.6). The cohort included 71 female and 131 male patients, yielding a female-to-male ratio of 0.54. The most common histopathology was undifferentiated pleomorphic sarcoma (UPS). 30% of the patients had high-grade tumors while 36 had intermediate grade tumors. Grading did not apply to 36% of the tumors. Additionally, 56 patients had viable tumor tissue on surgical margins. The only clinical variable that showed statistical significance between the case and control groups was lower extremity location of the tumor. Refer to **Table 4.1** for additional details about the characteristics of the cohort.

4.4.2 Inter-reader agreement:

Tumor volume segmentations by the two radiologists exhibited a high level of agreement, with a Dice similarity coefficient of 0.87 (SD = 0.08). Interobserver agreement was excellent, with an average value of 0.92. The radiologists had 100% agreement in identifying individual soft tissue masses.

4.4.3 Univariate Analysis

Adjusting for age, sex, tumor histopathology and grading and correcting for multiple testing, the only variable significantly associated with involvement of post-surgical margin status was intratumoral enhancement (adjusted p-value < 0.05).

4.4.4 Radiologist Evaluations Model Performance

The semantics-based model achieved an average AUROC of 0.72 in cross-validation while reaching an AUROC of 0.62 (0.45 – 0.79) on the test set and an AUROC of 0.75 (0.51 – 0.99) on the longitudinal test set. The most important features in this model were patient age, pretreatment tumor margin definition on the T2-weighted and contrast enhanced images and the size of the tumor before treatment. Feature importance's can be viewed in **Figure 4.5**.

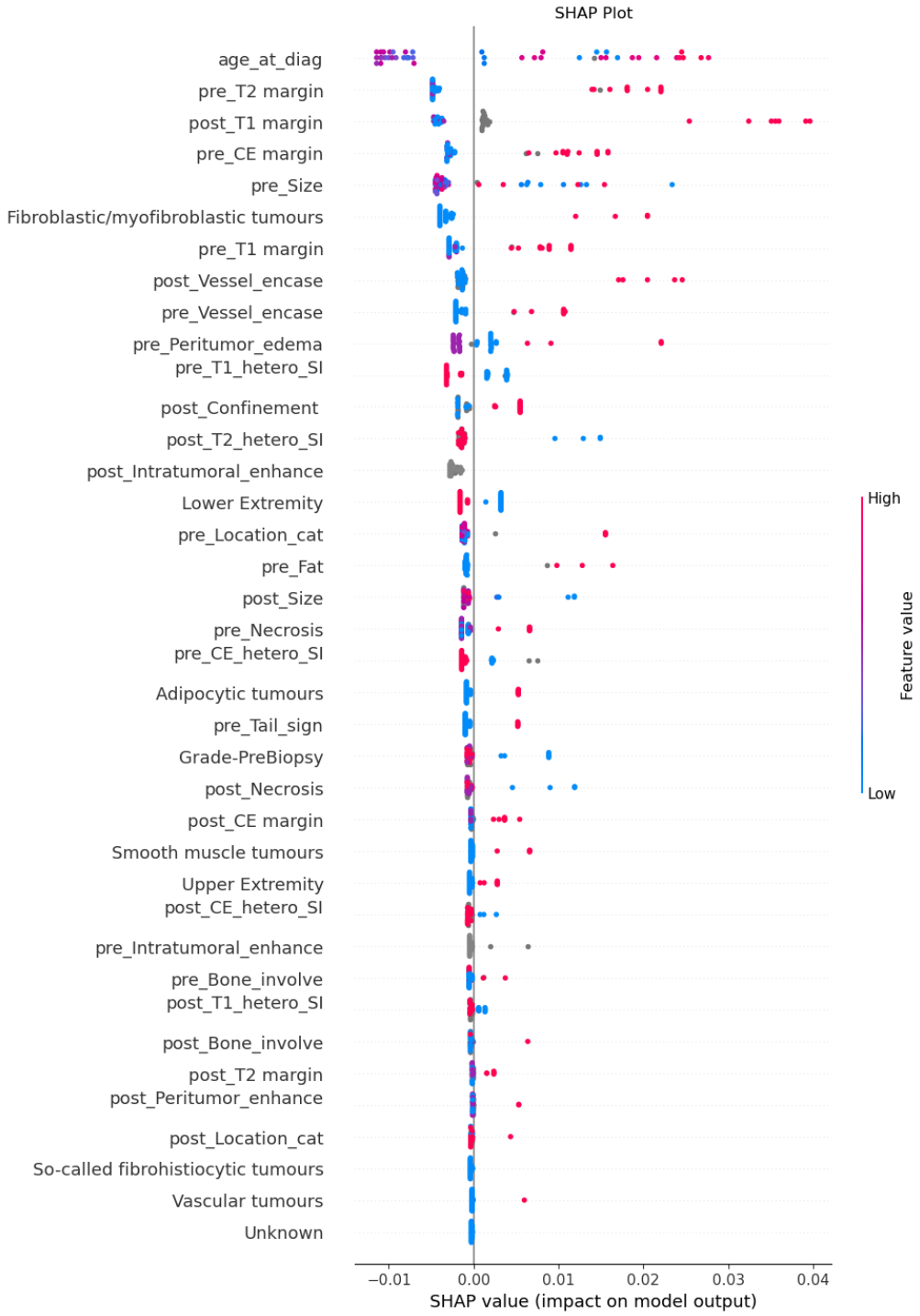

Figure 4.5 The SHAP feature importance plot for the semantics model. Features higher in the plot were more important to the model during prediction of the label in the test set. Each dot is one example from the test set and the color of each dot represents its value with red representing high

values and blue representing low values. In case of binary variables, red represents positive and blue represents negative. Dots on the left side of the central line mean that that specific value nudged the model towards making a negative prediction (clear margins) and to the right of the central line mean that the value nudged the model toward making a positive prediction.

4.4.5 Unimodal Model Performance

The clinical model achieved an AUROC of 0.55 (0.35 – 0.75) on the independent test set and an average AUROC of 0.66 on cross-validation. It achieved an AUROC of 0.74 (0.60-0.89) on the longitudinal test set. The most important clinical features were age at the time of diagnosis, tumor grading and the fibroblastic/myofibroblastic subtype.

The imaging only model using only single time-point pre-treatment images achieved an AUROC of 0.74 (0.60 – 0.89) on the test set while achieving an average AUROC of 0.73 on cross-validation. On the longitudinal test set, it achieved an AUROC of 0.83 (0.67 – 0. 98). By applying the optimum threshold for prediction using the Youden index, it achieved a sensitivity of 43% and a specificity of 67%.

4.4.6 Multimodal Model Performance

Among the multimodal models, the intermediate fusion model was the best overall performing model achieving an average AUROC of 0.76 on cross-validation. It achieved a AUROC of 0.80 on the test set, and an AUROC of 0.82 (0.61 – 0.99) on the longitudinal test set. Using optimum thresholds for prediction, it achieved a sensitivity of 0.67 and specificity of 0.78.

The early fusion method achieved an average AUROC of 0.70 in cross-validation while reaching an AUROC of 0.74 (0.58 – 0.89) on the test set. On the longitudinal-only subset of the test set, it reached an AUROC of 0.79 (0.56 – 0.99). The late fusion model achieved an AUROC of 0.80 on

cross-validation, 0.72 (0.56 – 0.87) on the test set and 0.74 (0.66 – 0. 83) on the train set. Refer to **Table 4.3** for comparison of the different models and additional performance metrics. Confusion matrices and ROC plots can be seen in **Figures 4.6-4.8**.

Table 4.3: Comparison of unimodal and multimodal model performance metrics.

Model Name	Average Cross-Validation AUROC	Test AUROC (95% Confidence Interval)	Test AUROC on Longitudinal (95% Confidence Interval)	Train AUROC (95% Confidence Interval)	Test Sensitivity	Test Specificity
Clinical Model	0.66	0.55 (0.35 – 0.75)	0.76 (56 - 96)	0.74 (0.65 – 0.83)	50%	79%
Deep Learning Imaging	0.73	0.74 (0.60 – 0.89)	0.83 (0.67 – 0.98)	0.70 (0.60 – 0.79)	43%	72%
Semantics Model	0.72	0.62 (0.45 – 0.79)	0.75 (0.50 – 0.99)	0.82	43%	67%
Multimodal Deep Learning Models						
Early Fusion	0.70	0.74 (0.58 – 0.89)	0.79 (0.56 – 0.99)	0.66 (0.54 – 0. 75)	42%	83%
Intermediate Fusion	0.76	0.80 (0.66 – 0.95)	0.82 (0.61 – 1.0)	0.75 (0.65 – 0.85)	67%	78%
Late Fusion	0.80	0.72 (0.56 – 0.87)	0.79 (0.62 – 0.97)	0.74 (0.66 – 0.83)	33%	90%
AUROC: Area under the receiver operator curve						

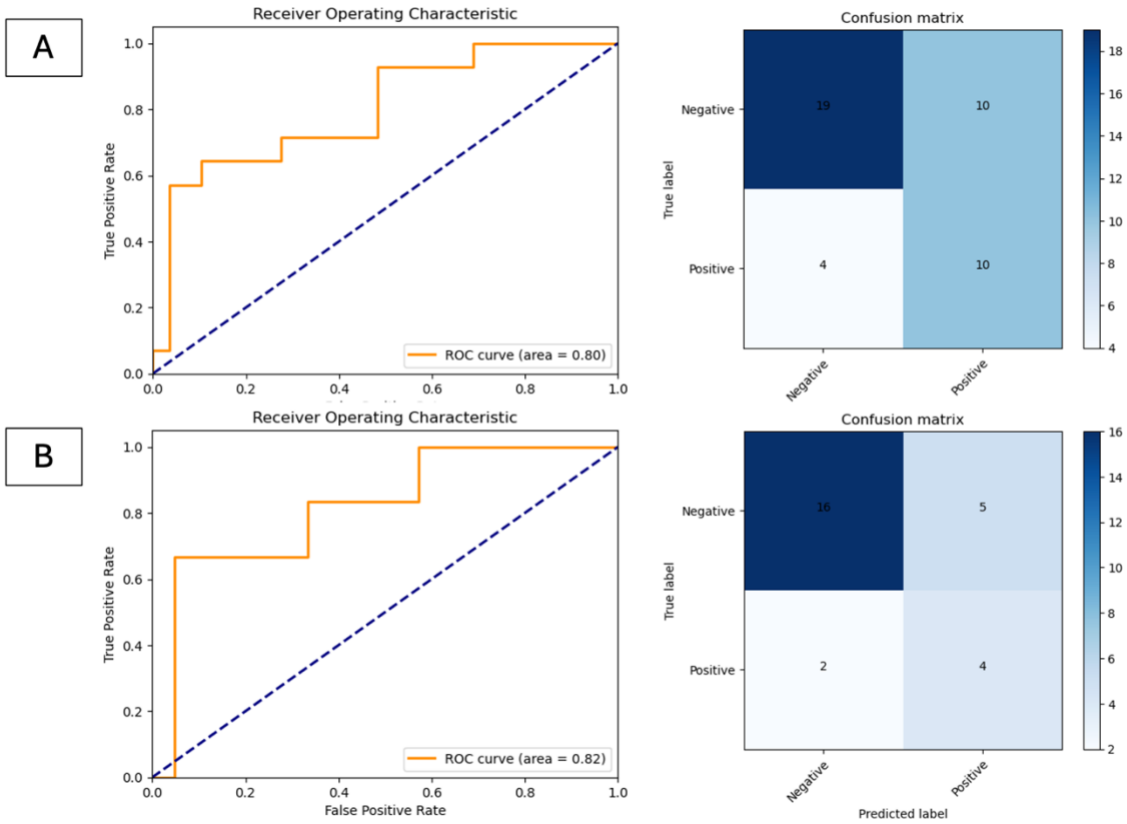
A
Receiver Operating Characteristic
True Positive Rate
ROC curve (area = 0.80)
Confusion matrix
Negative
Positive
True label
19
10
4
10
B
Receiver Operating Characteristic
True Positive Rate
False Positive Rate
ROC curve (area = 0.82)
Confusion matrix
Negative
Positive
True label
16
5
2
4
Predicted label

Figure 4.6: Performance plots of the intermediate fusion model. The receiver operator curve (ROC) can be viewed on the left and the confusion matrix can be viewed on the right. A. Performance over the entire test set. B. Performance over the longitudinal section of the test set.

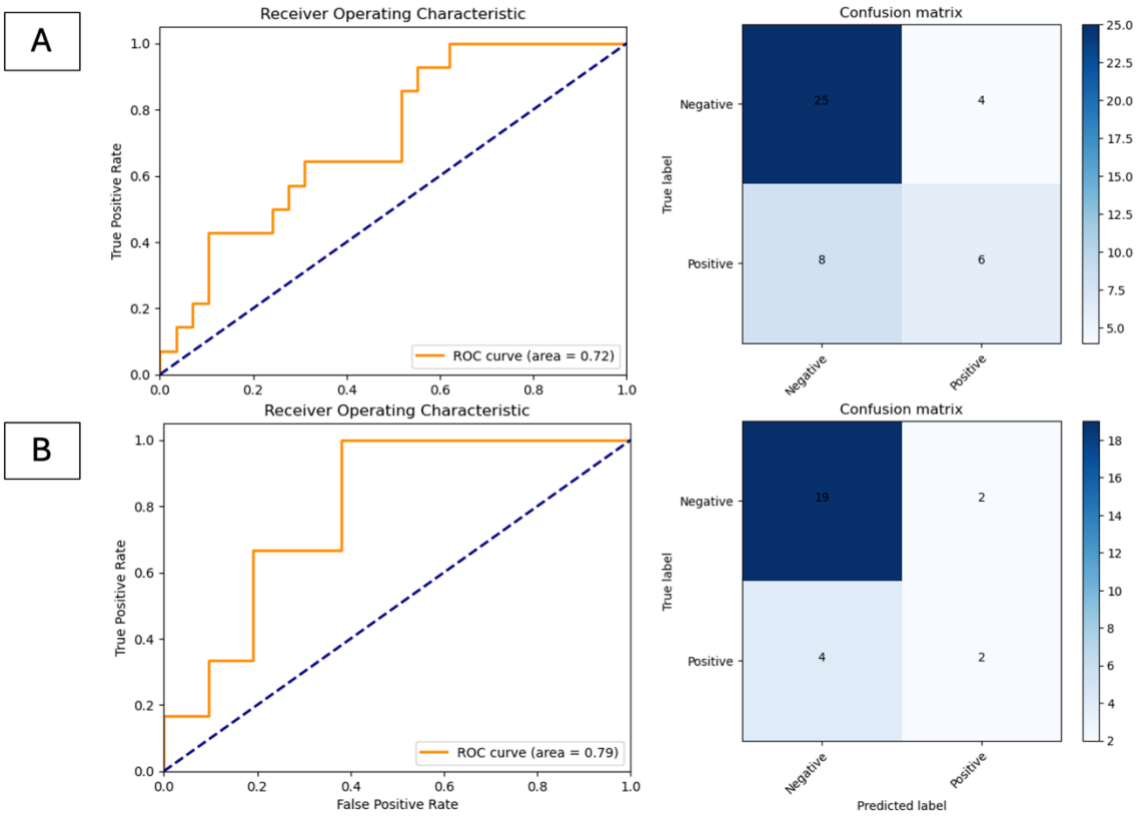

Figure 4.7: Performance plots of the early fusion model. The receiver operator curve (ROC) can be viewed on the left and the confusion matrix can be viewed on the right. A. Performance over the entire test set. B. Performance over the longitudinal section of the test set.

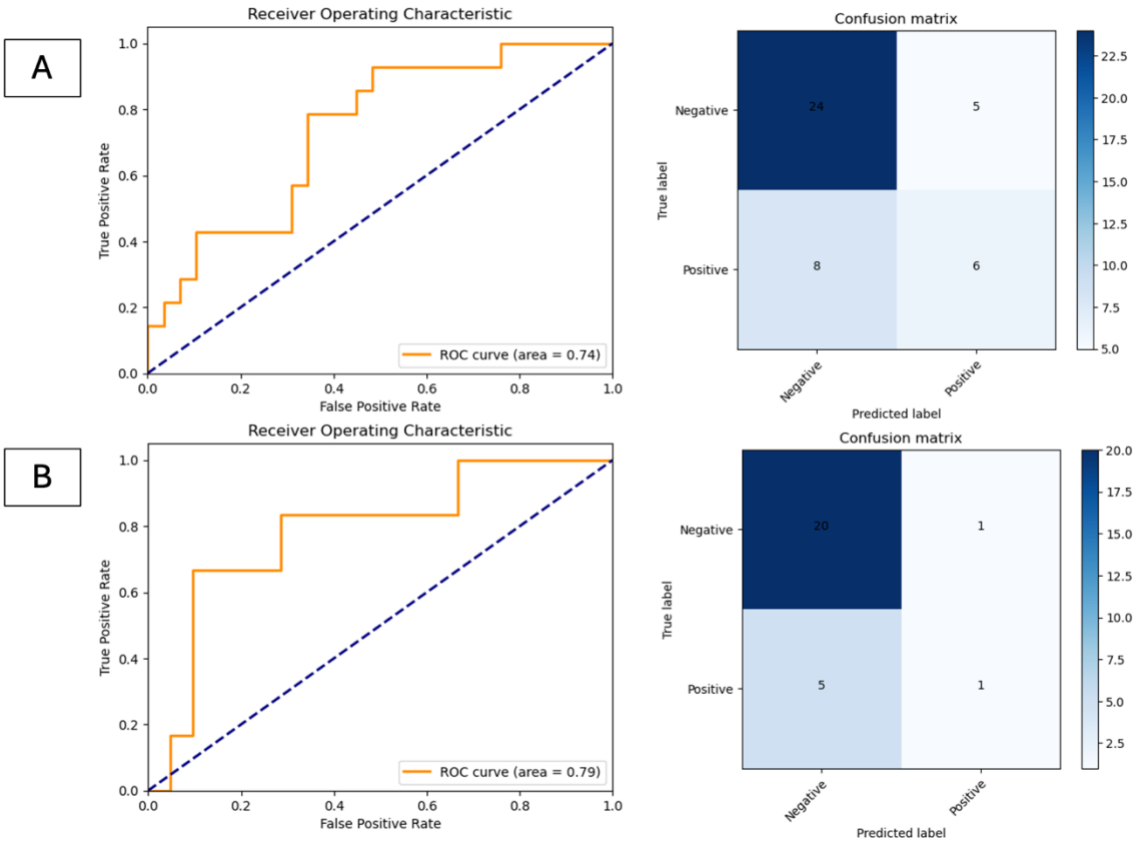

Figure 4.8: Performance plots of the early fusion model. The receiver operator curve (ROC) can be viewed on the left and the confusion matrix can be viewed on the right. A. Performance over the entire test set. B. Performance over the longitudinal section of the test set.

4.5 Discussion:

This aim focused on development and assessment of the incremental value of combining a deep convolutional neural network feature extractor on imaging data and clinical data on a binary

prediction task: Predict post-surgical margin status in soft tissue sarcoma. Given my overarching question of evaluating multimodal fusion strategies in predictive clinical model, here I focused on integration of 3D imaging and clinical data for a binary outcome prediction task. In this aim, we demonstrated strategies to use early, intermediate, and late fusion approaches in combining imaging and clinical data. In addition, we showed that intermediate fusion strategy had higher performance compared to early and late fusion strategies. Our study highlights the ability to screen patients who are at a high risk of having involved post-surgical excision margins in patients with soft tissue sarcoma. In addition, we demonstrated how using longitudinal MRIs and data fusion strategies to incorporate multimodal information improves the predictive performance of the predictive models. Our models have the potential to improve surgical planning by highlighting high risk patients and providing explanations on why the patient was selected as a high-risk patient.

Although post-surgical margin status is one of the most important predictors of disease recurrence, there are no established methods to estimate risk of involvement of excision margins (64). Such criteria exist for other outcomes like the modified CHOI and RECIST criteria which are used to estimate treatment response (72,73). Some studies have shown that factors like pre-tumoral edema, existence of a tail sign or lack of well-delineated tumor borders in MRI as signs of tumor invasion into the surrounding tissue may be helpful (74). Machine learning has been extensively used to predict tumor characteristics and disease outcome in patients with STS (65,75–79). These include tumor grading, treatment response to neoadjuvant therapy, disease-free survival, and overall survival. Despite this, few studies have investigated prediction of post-surgical margin status. Notably, a study by Lee et al. have investigated using radiomic features extracted from T2 DIXON sequences to predict post-surgical margin status and compared the performance of their model with radiologists in a small cohort of patients (69). Their work shows the promise of using quantitative

MRI measures in predicting post-surgical margin in soft tissue sarcoma. However, they use the DIXON sequence that may not be available in many centers. In addition, they only utilize single timepoint images. Our study builds on top of the previous work by incorporating multiple new components that have been shown to improve performance in other domains. First, we use multi-parametric MRI using information from multiple sequences. We also use longitudinal MRIs where available. Changes in the tumor characteristic during neo-adjuvant treatment may provide additional information regarding the characteristics of the tumor that are relevant to margin prediction. By combining clinical, pathologic, and imaging data in our deep learning model, we provide a more comprehensive view of the disease. This is especially important in soft tissue sarcoma in which tumors can have diverse histopathologies. By incorporating this information into the model, we allow the model to make histology specific decisions. Utilization of explainability tools allows for case specific and aggregate explanations that can shed light on the reasoning behind the model's predictions. Additionally, we compare our deep learning model performance with semantics-based models that are derived from radiologist assessment to identify any added value from using deep learning.

In this study, we experimented with three established approaches of multimodal data fusion in deep learning. While each of the early, intermediate, and late fusion strategies offer advantages and disadvantages, it is unclear what fusion strategy would work in a specific scenario (10). Many studies just stick to one fusion strategy without explanation of why that strategy was chosen and some treat it as a hyperparameter. In this study we explored the implications of using each of those fusion strategies in an example question, hoping to shed some light on some of the benefits and downsides of using each. Interestingly, based on various outcome measures, different fusion

strategies can be selected as the best performing. However, it is noteworthy that due to the sample size of this study, some of the differences in performance are not statistically significant.

In general, if all train, cross-validation, and test performance are taken into consideration together, the intermediate fusion strategy was the best performing model. The intermediate and late fusion strategies outperformed early fusion in prediction of margin status in both test sets. Similar studies have come to similar conclusions in the past (24,80). The lower performance of the early fusion model can be attributed to the fact that the early fusion model only utilizes features extracted from a pretrained ResNet model and does not finetune the model based on the imaging studies in our data. This can be beneficial when the number of samples is very small but given our sample size, our study shows that some finetuning improves the model performance.

4.5.1 Limitations

Some limitations of our study included the diversity of STS, which made subtype-specific modeling difficult. By incorporating subtype as a variable in the model, we aimed to make the model aware of the different subtypes and address this gap. Additionally, not all of our cases had longitudinal data. We addressed this problem by including those cases in the training process so that we don't use valuable information. We designed our model in a way that allowed it to deal with missing sequences and missing timepoints.

4.5.2 Conclusions

This aim was focused on extending our understanding of multimodal data fusion approaches in predictive models by investigating the implications of various fusion strategies in combining imaging and clinical data in a binary prediction task. We demonstrated the added value of using multi-timepoint, multimodal deep learning with proper fusion strategy in prediction of post-

surgical margin involvement in soft tissue sarcoma. Our model showed superior performance when compared to a model developed based on radiologist evaluations. In addition, our best performing model achieved the best performance when longitudinal data was provided to the model. Our models can be used in clinical practice to identify patients that are at a higher risk of involvement of post-surgical margins. Furthermore, I showed how using the intermediate fusion strategy outperformed other strategies closely followed by late fusion in combining volumetric imaging data with structured clinical data. Future directions could involve the assessment of the utility of this risk score in reducing positive margin incidence in patients with soft tissue sarcoma. Over the last two chapters, I have looked at various facets of my overarching question. In Aim 1, I looked at combining longitudinal EHR, genomics and survey data. Meanwhile in Aim 2, I explored combining 3D imaging and clinical variables. While both aims were focused on a binary outcome prediction task, in the next chapter, I will explore multimodal data fusion of 2D imaging and clinical data for a continuous regression task. Lessons learned in this aim about fusion can be seen in **Figure 4.9**.

Early Fusion	Intermediate Fusion	Late Fusion
• Good choice if features can be represented with no simplification/dimensionality reduction. • Can handle missing data if appropriate architecture is used. • **May lose some signal when feature extractors are used.** • **May help when sample size is very small for finetuning on imaging.**	• May overfit to training data due to large number of trainable parameters. • **Better for combination of imaging and clinical data.** • **Various network parts train at different speeds and may need different hyperparameters. Pretraining helps with stable training.**	• Can combine diverse networks like XGBoost and transformer. • **Does not learn inter-modal correlations.** • **Less overfitting due to lower number of trainable parameters.** • **Performance depends on unimodal models.**

Figure 4.9 Lessons learned about fusion in Aims 1 and 2.

5. Aim 3: Evaluation and comparison of early, intermediate, and late fusion techniques for combining imaging and clinical data on a regression prediction task: Estimation of CT-based body composition metrics from chest radiographs.

This chapter is a modified and extended version of a paper titled "XComposition: Multimodal Deep Learning Model to Estimate CT-based Body Composition Measures Using Chest Radiographs and Clinical Data" authored by Ehsan Alipour, Peter Tarczy-Hornoch, MD, Jennifer Hadlock, MD, Nickolas Stucky, MD, PhD, et al., originally published as a preprint at https://doi.org/10.1101/2025.01.16.25320684 and submitted to Radiology: Artificial Intelligence on March 2, 2025. Text in italics is directly drawn from the original manuscript. Additions and revisions are presented in standard font.

In Chapter 2, I introduced my overarching question as evaluating multimodal data fusion approaches for clinical predictive models. In the course of the last two aims, I explored the application of early, intermediate and late fusion strategies using multiple medical data modalities. Notably, I explored fusion of EHR data with genomic and survey data in Aim 1 while introducing a novel transformer model that incorporates the entirety of the structured clinical information and takes into consideration time differences and measurements/laboratory test results when applicable. In this aim, the outcome of interest was a binary prediction task of whether a patient will progress to CKD in the next 5 years after their initial diagnosis of type 2 diabetes. In Aim 2, I extended the concept of multimodal data fusion to the imaging domain by attempting to combine cross-sectional, multiparametric imaging data with select clinical and pathologic features extracted from the electronic health records. Similar to the previous aim, aim 2 also focused on using multimodal data fusion for a binary prediction task. Aim 2 showed how intermediate fusion had a slight edge, closely followed by late fusion, we attempting to predict involvement of post-surgical

tissue margins in soft tissue sarcoma. In this chapter, I further explore the concept of multimodal data fusion when it comes to imaging and clinical data by using a different 2D imaging modality (Chest Radiographs) and using multimodal data fusion to estimate a series of continuous variables (Body composition metrics extracted from CT scans). Through this aim, I hope to generate more evidence for my findings in the previous aims that intermediate and late fusion strategies are the preferred methods of data fusion. Additionally, I will use a slightly different version of early fusion in this aim to address the limitation in Aim 2 that the choice of the pretrained feature extractor CNN might have contributed to the low performance. In this aim, instead of using a pretrained CNN to extract features from the imaging, I will use a small one-layer fully connected neural network to generate embedding from the clinical data that are then added to the imaging data. Using this approach, I hope that the CNN is able to fine tune on the imaging and clinical data together.

5.1 Background

Body composition metrics have attracted significant attention as predictors of health-related outcomes. Multiple studies have demonstrated correlations between body composition metrics like visceral fat area and skeletal muscle index and chronic non-communicable disease including cardiovascular disease and diabetes (81,82). In addition, studies have found that abnormal body composition measures can be a predictor of poor prognosis in patients with cancer including those with breast and colorectal cancers(83). These correlations are often stronger than correlation between these diseases and commonly used body composition surrogates like body mass index (BMI) and weight (81,82,84). Measures such as BMI fail to capture critical information about body composition including muscle mass, visceral adipose tissue (which has been shown to be the associated with pathologic conditions (85)) and subcutaneous adipose tissue. Studies have also

explored the use of body composition metrics acquired through CT scan or MRI for opportunistic screening of various diseases including cardiometabolic disease, osteoporosis, and steatohepatitis (86).

Common methods for calculating body composition metrics include whole body magnetic resonance, computed tomography imaging or dual-energy x-ray absorptiometry (DXA), and bioelectrical impedance analysis (87). However, these methods may not be available for all patients due to limitation of resources, risk of radiation exposure, or need for specialized equipment, software and staff. Scientists have validated using body composition metrics calculated from 2D slices in the L3 section of abdominal CT scans as surrogates of whole-body composition metrics (88,89). While CT and MRI provide accurate estimates of body composition (84), performing a CT scans or MRI to calculate body composition metrics on the general public for opportunistic screening poses challenges, including high costs and risk from higher radiation exposure(84,87).

5.2 Significance

As described in Chapter 2, there is a paucity of studies on exploring the implications of using different fusion strategies in combining imaging and clinical data. Also, many of the studies that use multimodal imaging and clinical data are focused on binary prediction tasks. In this chapter, I explore my overarching question further by exploring fusion of 2D imaging and clinical data in the context of a continuous regression task. Specifically, I will use chest radiographs and relevant clinical variables to estimate CT-based body composition metrics.

Given the limitations of the current body composition calculation methods, new strategies to estimate body composition measures using readily accessible clinical and imaging data would enable calculation of body composition metrics for a larger pool of individuals. One promising

avenue is to use simpler modalities like chest radiographs to estimate body composition. A recent study on opportunistic screening of type 2 diabetes demonstrated that chest radiographs encode information about fat distribution that can be used to screen for type 2 diabetes (90).

In this study, we aim to create a multi-modal deep learning model to estimate CT-based body composition metrics by combining readily available clinical data with chest radiographs. Such a model can be applied with minimal cost retrospectively or prospectively on any person with a chest radiograph, enabling greater use of body composition metrics in research, and clinical care including screening. In addition, we will experiment with different ways of combining the clinical data and imaging data. In the next section, I will go over the methods used to curate the dataset and train and evaluate the multimodal deep learning models.

5.3 Materials and Methods

To ensure that our study meets the requirements of high-quality clinical AI research, we followed the checklist of artificial intelligence in medical imaging (CLAIM) guidelines in all steps of our study (70).

5.3.1 Dataset

The clinical and imaging data used in our study was compiled from Truveta Data accessible from Truveta.com. Truveta provides access to continuously updated and linked EHR data from around 200 patients. It consisted of deidentified records from *the same* subset of *subjects selected for* imaging data. The data used in this study was accessed on July 14th, 2024. This study used only de-identified patient records and therefore did not require Institutional Review Board approval.

We identified adult individuals who had a chest radiograph within 3 months of an abdominal CT scan without any contrast agent. We further limited our inclusion criteria to patients with a weight

measurement within 1 month of the chest radiography and had a height measurement anytime in their adult life. Although our search returned 312,444 unique cases, due to computational constraints, we randomly sample 3,000 patients. The imaging data, in addition to the age at the time of imaging, sex at birth, weight, and height were collected for every patient. De-identification of the data was attested to through expert determination in accordance with the HIPAA Privacy Rule prior to our data retrieval. *For* patients with multiple imaging that fit the criteria, the last abdominal CT and the closest posteroanterior (PA) or anteroposterior (AP) chest radiography to the abdominal CT were selected. Weight was determined from the observation closest to the date of the radiography. In addition to height and weight, we determined the age at time of chest radiography and the sex at birth for all patients. We normalized all weights to kilograms and all heights to meters. The cohort flowchart diagram can be seen in **Figure 5.1.**

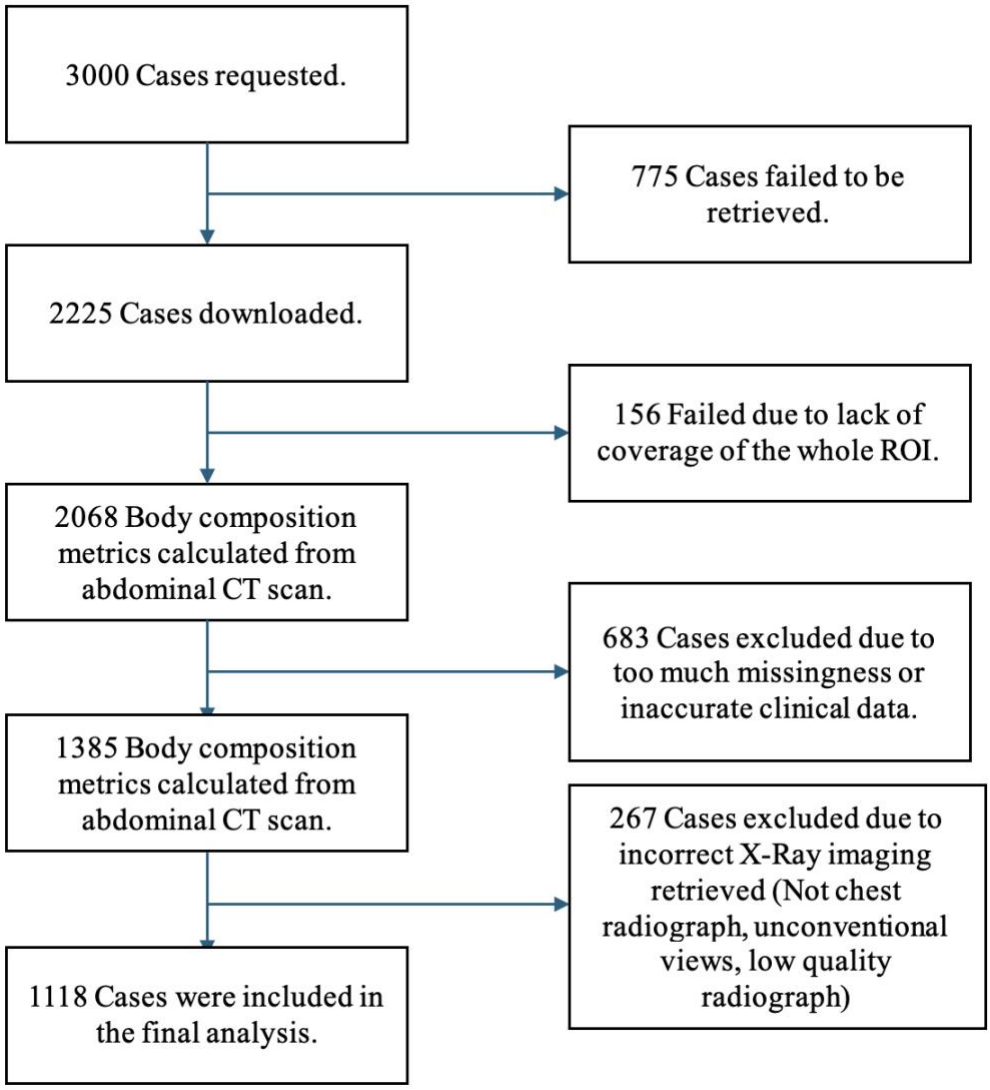

Figure 5.1 Cohort Flowchart. We initially request 3000 cases. Out of these, 1118 cases met all our inclusion criteria and were included in our study. ROI: Region of Interest.

5.3.2 Missing Data

For cases that the recorded height or weight were outside of acceptable range or were missing, we imputed the missing value using linear regression on the rest of the variables.

5.3.3 Body Composition Metrics Calculation

The ground truth values for body composition were calculated using a combination of 2D slices from the T12 vertebrae to the L5 vertebrae for volumetric body composition metrics and on the mid-L3 level for the single slice body composition metrics in the abdominal CT scans. The TotalSegmentator (91) tool was used to segment out the various body composition sections in CT scans and identify the T12, L3 and L5 levels. The list of body composition metrics that were calculated can be found in **Tables 5.1** and **5.2**. An in house rule based python script based on available literature was used to calculate the body composition metrics from the segmentation masks (92–94). Area was calculated by multiplying the number of pixels in each region of interest by the area of each axial pixel. Calcified plaques were identified using any contiguous voxels with a Hounsfield units (HU) value significantly larger than the median HU value of the aorta as described in the paper (13). Agatston scoring algorithm is then used to score the abdominal aorta calcification (14). Skeletal muscle fat volume was calculated by multiplying pixel area by the number of pixels in the skeletal muscle with a HU value in the range of adipose tissue (15). For abdominal aorta calcification score and number of calcified plaques only volumetric segmentation of abdominal aorta was used.

Table 5.1 Population Characteristics and Volume Based Body Composition Distribution Across the Training, Validation and Test sets.

		Train (n = 726)		Validation (n = 177)		Test (n = 215)		All (n = 1118)		p-value
Variable	Category	mean	std	mean	std	mean	std	mean	std	
Age		67.13	17.36	68.15	16.31	66.13	16.8	67.1	17.09	0.51
Height (Meters)		1.69	0.18	1.66	0.16	1.7	0.27	1.69	0.2	0.12

Weight (Kilograms)		77.34	19.66	79.66	22.11	77.74	20.28	77.78	20.18	0.39
		Count	%	Count	%	Count	%	Count	%	
Sex	Female	374	51.52	90	50.85	118	54.88	582	52.06	0.64
	Male	352	48.48	87	49.15	97	45.12	536	47.94	
		mean	std	mean	std	mean	std	mean	std	p-value
Skeletal Muscle Volume (cm^3)		2241.42	783.94	2265.02	848.56	2233.83	843.84	2243.69	805.52	0.92
Visceral Fat Volume (cm^3)		2850.55	1940.01	2877.18	1734.09	2791.08	1916.73	2843.33	1902.95	0.89
Subcutaneous Fat Volume (cm^3)		4464.93	2832.39	4789.71	2782.95	4560.83	2700.3	4534.79	2799.63	0.38
Vertebral Bone Volume (cm^3)		347.09	76.06	347.6	85.02	342.74	73.76	346.34	77.06	0.75
Fat Free Volume (cm^3)		7982.08	2166.6	8249.24	2807.68	8137.64	2421.02	8054.29	2328.64	0.33
Intramuscular Fat (cm^3)		220.17	167.36	213.24	156.64	224.87	175.39	219.98	167.18	0.79
Vertebral Bone Density (HU)		320.87	78.46	322.87	71.53	329.11	86.03	322.77	78.93	0.4
Muscle Radiodensity (HU)		19.49	15.4	19.52	13.29	19.23	15.14	19.44	15.02	0.97
Aortic Calcification Score		8.67	14.58	9.22	13.97	8.72	12.55	8.77	14.11	0.9
Number of Plaques in Abdominal Aorta	10.62	11.16	10.17	10.67	10.64	11.06	10.55	11.05	0.88	

Skeletal Muscle Index	792.43	276.83	828.56	326.89	788.1	303.19	797.32	290.48	0.29
Subcutaneous Fat Index	1620.39	1099.3	1803.23	1198.27	1636.58	1013.87	1652.45	1100.84	0.14
Visceral Fat Index	999.76	667.44	1038.11	588.53	984.25	664.47	1002.85	654.65	0.7
Fat Free Index	2830.93	765.86	3026.78	1080.12	2888.8	906.78	2873.07	852.64	0.02

Table 5.2 *Distribution of mid-l3 level body composition metrics.*

	Train mean	Train std	Validation mean	Validation std	Test mean	Test std	All mean	All std	p-value
Skeletal Muscle Area (cm^2)	125.19	39.14	125.55	40.69	125.09	40.08	125.23	39.54	0.99
Skeletal Muscle Index	44.55	14.6	46.31	17.58	44.45	16.51	44.81	15.49	0.37
Visceral Fat Area (cm^2)	166.41	112.71	171.21	102.83	166.66	111.93	167.23	110.96	0.87
Visceral Fat Index	58.81	39.98	62.06	35.12	59.11	40.12	59.39	39.26	0.61

Subcutaneous Fat Area (cm^2)	238.02	149.96	255.73	142.55	237.12	138.73	240.67	146.71	0.33
Subcutaneous Fat Index	86.92	59.24	96.6	62.28	85.23	52.19	88.13	58.52	0.1
Fat Free Area (cm^2)	369.48	96.71	379.41	124.86	377.9	111.37	372.69	104.55	0.38
Fat Free Index	132.08	37.74	140.52	53.04	134.85	45.96	133.96	42.24	0.05
Intramuscular Fat Area (cm^2)	12.56	10.08	11.79	9.56	12.86	10.63	12.5	10.1	0.55
Vertebral Bone Area (cm^2)	18.93	4.84	18.98	4.91	18.51	5.18	18.86	4.91	0.51
Vertebral Bone Density (HU)	326.38	83.42	329.01	74.68	333.66	88.05	328.21	82.99	0.52

The data was split with a 0.8/0.2 ratio to generate the training dataset and the hold-out test set. The test set was only used to measure the final model performance. Additionally, 20% of the training data was used as a validation set for hyperparameter tuning.

5.3.4 Preprocessing and Normalization

All retrieved radiographs were manually reviewed by a physician scientist to ensure only PA or AP chest x-rays are included and images contain no significant artifacts. Chest radiograph pixel values were normalized using z-score normalization. All x-rays were resized to 512*512 pixels. Z-score normalization was applied on all numerical variables and all calculated body composition metrics to ensure numerical stability during training using mean and standard deviation of the training set.

5.3.5 Modeling Approach and Fusion Strategy

The clinical baseline model was developed using a shallow neural network with two fully connected layers. We also developed a regression model using the four clinical variables (height, weight, age, and sex at birth).

The imaging-only model consisted of a multitask convolutional neural network (CNN). Multiple CNN architectures (ResNet10, ResNet50, ResNet101) were tested, however ResNet18 (95) was selected based on validation performance. The network was initiated randomly. The default resnet18 architecture from the PyTorch (v2.4) package was used.

We experimented with three fusion strategies, comparing early, intermediate and late fusion of clinical and imaging data (9,96). **Figures 5.2-5.5** depict all fusion strategies. *For early fusion, we used a fully connected layer to generate encodings for the clinical variables that were in turn added to the input image and passed through the rest of the network. For intermediate fusion, the encodings generated by the CNN were concatenated with the encodings generated by the shallow neural network and a final fully connected layer was added on the top. For late fusion, two separate imaging-only and clinical-only models were developed. The predictions of these two networks were collected and passed through a fully connected neural network.*

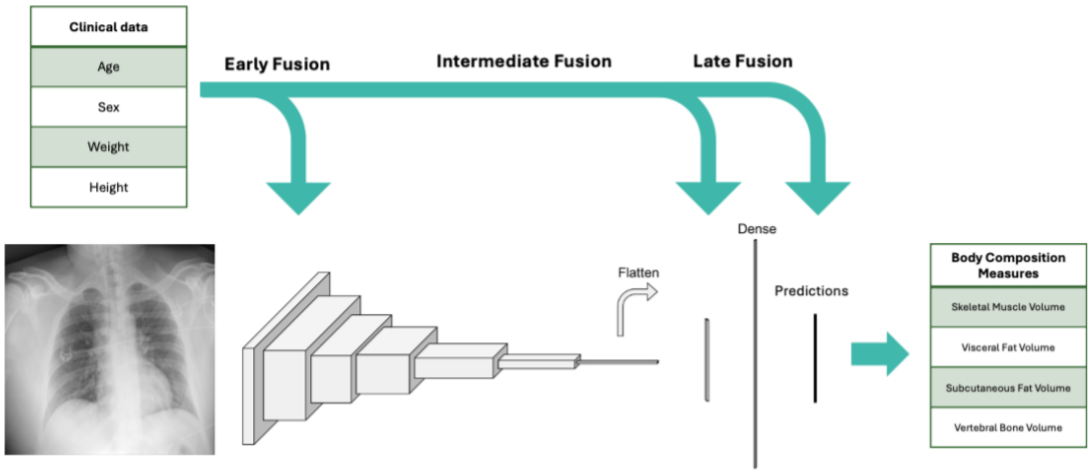

Figure 5.2 A diagram showing model architecture, input features, and the various fusion strategies used in our study. We used a resnet18 model for our convolutional neural network.

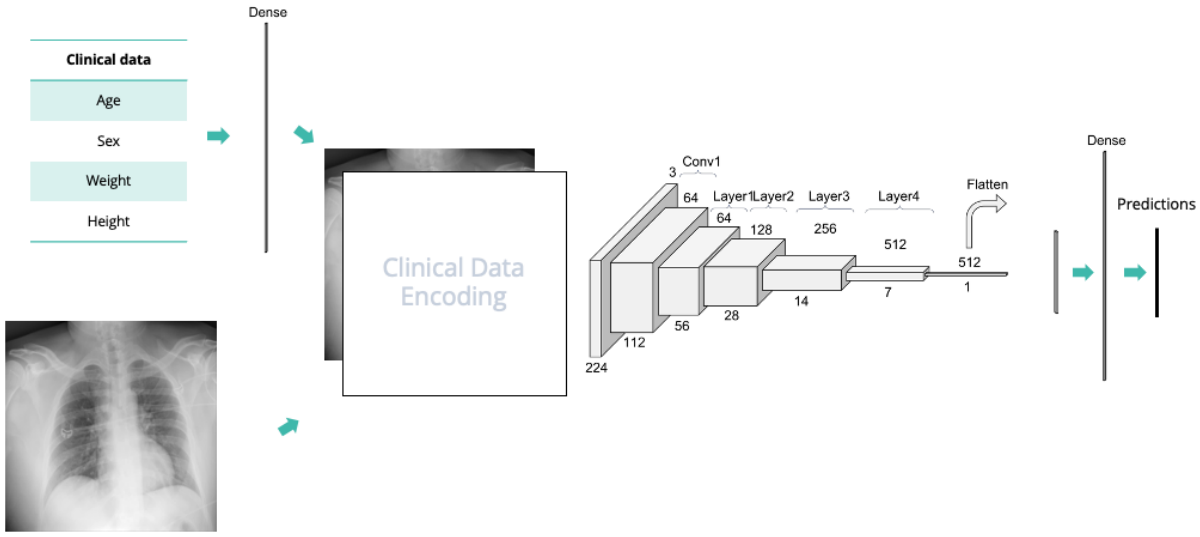

Figure 5.3 Early fusion strategy, in this network, a small fully connected layer is used to generate embeddings from the clinical data that are added to the imaging data before any further processing.

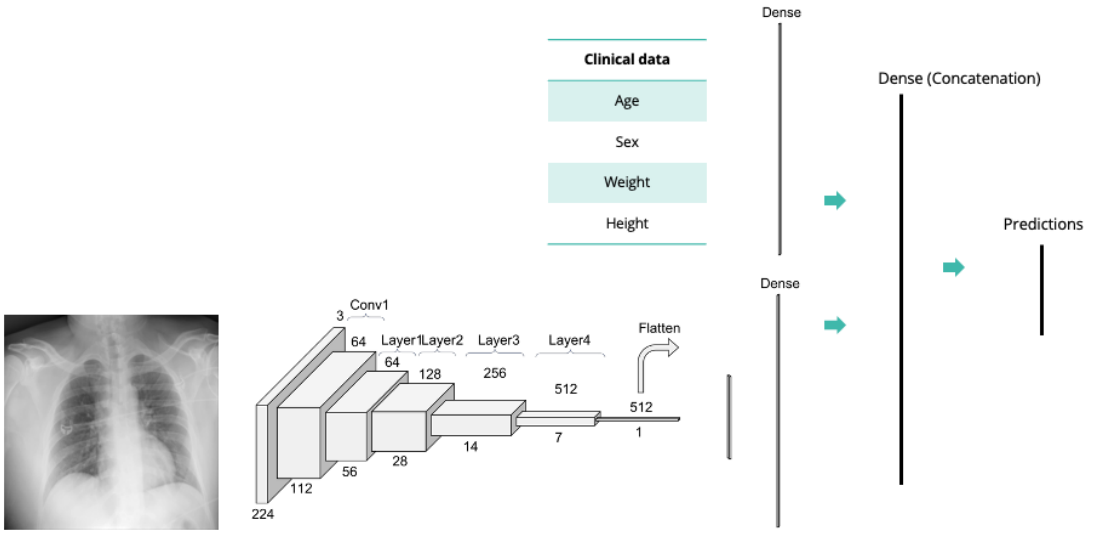

Figure 5.4 Intermediate fusion strategy. Embeddings are generated using a shallow neural network from the clinical data and the ResNet18 architecture from the imaging data. The embeddings are concatenated together, and a final neural network will be used to make predictions.

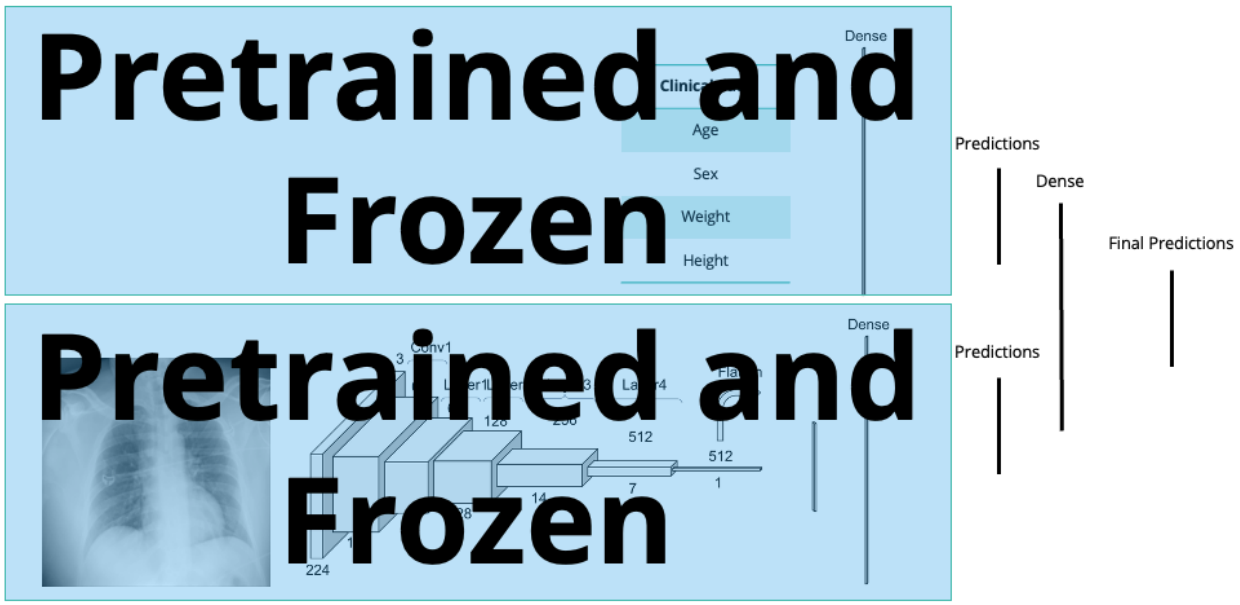

Figure 5.5 Late fusion strategy. In late fusion, two separate clinical and imaging networks are trained. The predictions that these networks make are used by another shallow neural network to make the final predictions.

The early fusion strategy used in this aim differs from the one used in Aim 2 in that in contrast to approaches that rely on a pretrained convolutional neural network (CNN) to extract features solely from imaging data, this method uses a lightweight, single-layer fully connected neural network to generate embeddings from the clinical data. These clinical embeddings are subsequently integrated with the imaging data prior to input into the CNN. The objective of this design is to enable the CNN to jointly fine-tune on both imaging and clinical features during training, rather than treating the two modalities separately or relying on fixed feature representations. By allowing end-to-end optimization of the combined input, this approach aims to enhance the model's ability to capture complex, multimodal relationships that may be critical for downstream predictive performance.

5.3.6 Evaluation

Huber loss (97), which is a combination of mean absolute error (MAE) and mean squared error (MSE) was used to train the models and evaluate the performance. In addition, weight decay was used for regularization. The correlation between the predicted outcomes versus the ground truth values is reported for all body composition metrics.

5.3.7 Explainability

Both occlusion sensitivity and integrated gradient methods were used to calculate saliency maps over all the images (98). Aggregate explainability maps were generated by averaging all the input images in the test set and their respective explainability maps for the two methods separately.

5.3.8 Fairness

Final model performance was measured and compared across the different age, sex and BMI groups present in our test set to calculate performance metrics in different subgroups and identify potential fairness issues with the model.

5.3.9 Statistical Analysis

P-value significance level threshold was set to 0.05 for all analyses. Pearson correlation was calculated to compare predictions and ground truth values for all body composition metrics except aortic calcification score and number of calcified plaques. Spearman correlation was used for those due to their skewed distribution. ANOVA was used to identify any statistically significant difference present across the data splits for continuous variables and chi square test was used for categorical variables. The following cut points were defined for the correlations: $r<0.6$ poor, $0.5<r<0.7$ moderate, $0.7<r<0.8$ good, $0.8<r<0.9$ very good and $r>0.9$ was considered great correlation.

5.3.10 Data sharing and Code Availability

The data used in this study is available to all Truveta subscribers and may be accessed at studio.truveta.com. Our model weights and the required code to calculate body composition metrics is shared on GitHub at https://github.com/Truveta/xcomposition_multimodal_deep_learning_body_composition. This project was done using PyTorch (v 2.4), Sci-kit Learn (v. 1.5.2), Captum (0.7.0) on GPU equipped Linux Azure server with one V100 GPU with 16 gigabytes of memory.

5.4 Results

5.4.1 Study Population

From a potential cohort of 3000 cases, we included a final cohort that consisted of 1118 subjects. Some cases were excluded due to problems with identifying the vertebrae levels. Other problems included missingness of both weight and height values, and low-quality chest radiographs, including the image not being a chest radiograph (mislabeled DICOM description), unconventional views, low exposure, or presence of large artifacts in the CXR. The cohort flowchart is shown in **Figure 5.1**. Average age was 68 ± 17.36 years of age. The average BMI was 27. The demographic distribution of the cases and distribution of the body composition metrics can be seen in **Table 5.1**.

5.4.2 Clinical Model

The clinical model included age, sex at birth, height, and weight. The clinical only model achieved good correlation (0.77, 0.71 – 0.82) when predicting subcutaneous fat volume and index but achieved moderate to poor performance in estimating other body composition metrics. The other highest performing metrics included visceral fat area 0.69 (0.65 – 0.71) and vertebral bone volume with a correlation of 0.67 (0.59 – 0.74). More detailed performance metrics can be viewed in **Table 5.3**.

5.4.3 Imaging Model

The image only model outperformed the clinical model in prediction of subcutaneous fat volume (0.78, 0.72 – 0.82) and visceral fat volume (0.70, CI 0.63 - 0.76) achieving good correlation with ground truth in both metrics. The imaging only model also performed better in estimating muscle radiodensity and vertebral bone radiodensity compared to the clinical-only model.

5.4.4 Combined Model

Our multimodal models achieved higher performance in estimating almost all body composition metrics, especially subcutaneous and visceral fat related measures. Our results showed that late fusion based model performed best in estimating body composition metrics followed closely by intermediate fusion. Scatter plots comparing model performance on the test set for the three top performing body composition metrics could be viewed in **Figure 5.6**. **Table 5.4** contains the final multimodal model performance metrics across the train, validation, and test cohorts.

5.4.4.1 Early Fusion

The early fusion model estimates achieved good correlation for subcutaneous fat index (0.80, 0.75 - 84) and achieved good correlation for visceral fat index, visceral fat volume and subcutaneous fat volume. The early fusion model consistently underperformed the two other models in estimating all body composition metrics.

5.4.4.2 Intermediate Fusion

The intermediate fusion model achieved a correlation of 0.82 (0.78 – 0.86) in estimating subcutaneous fat volume and index. It also achieved good performance in estimating visceral fat volume and visceral fat index. This model outperformed the other fusion strategies in estimating muscle radiodensity (0.69, 0.61 – 0.75), fat free index (0.60, 0.51 – 0.68), intramuscular fat (0.57, 0.48 – 0.66), and vertebral bone radiodensity (0.49, 0.38 – 0.58).

5.4.4.3 Late Fusion

The late fusion model achieved a correlation of 0.85 (0.81 - 88) in estimating subcutaneous fat volume. This model also had good performance in estimating visceral fat volume (0.76, 0.69 – 0.81) and vertebral bone volume (0.72, 0.65 – 0.78). The Spearman's correlation for skeletal

muscle volume (r = 0.67), vertebral bone volume (r = 0.75) and subcutaneous fat index (r = 0.88) was higher than the Pearson's correlations highlighting a slightly better ability to rank observations compared to estimating their actual value. Table 5.5 contains performance metrics for all body composition measures across the fusion strategies.

Table 5.3, Final model performance in the holdout test set comparing imaging only, clinical only and multimodal models. Values in the table are Pearson's correlations between predictions and ground-truth values.

	Skeletal Muscle Volume	**Visceral Fat Volume**	**Subcutaneous Fat Volume**	**Muscle Radiodensity**	**Fat Free Volume**
Imaging	0.45	0.7	0.78	0.61	0.47
Clinical	0.58	0.69	0.77	0.6	0.56
Multimodal	0.58 (95% CI: 0.49 - 0.67)	0.76 (0.65 - 0. 80)	0.85 (0.81 - 0.88)	0.69 (0.61 - 0.75)	0.59 (0.50 - 0.67)
	Intramuscular Fat	Aortic Calcification Score (Spearman)	Number of Plaques in Aorta (Spearman)	Skeletal Muscle Index	Visceral Fat Index
Imaging	0.48*	0.44	0.45	0.26	0.6
Clinical	0.52	0.69	0.67	0.51	0.67
Multimodal	0.55** (0.44 - 0.63)***	0.66	0.68	0.53 (0.43 - 0.62)	0.75 (0.67 - 0.80)
	Subcutaneous Fat Index	Fat Free Index	Vertebral Bone Volume	Vertebral Bone Density	
Imaging	0.74	0.31	0.5	0.47	
Clinical	0.77	0.56	0.67	0.38	
Multimodal	0.85 (0.80 - 0.88)	0.59 (0.49 - 0.67)	0.72 (0.65 - 0.78)	0.48 (0.36 - 0.57)	

*** Pearson Correlation between predictions and ground truth values reported for all models unless stated otherwise.**
**** Under each metric, the best performing model is shown with bold letters.**
***** The 95% confidence interval is reported for the test set.**

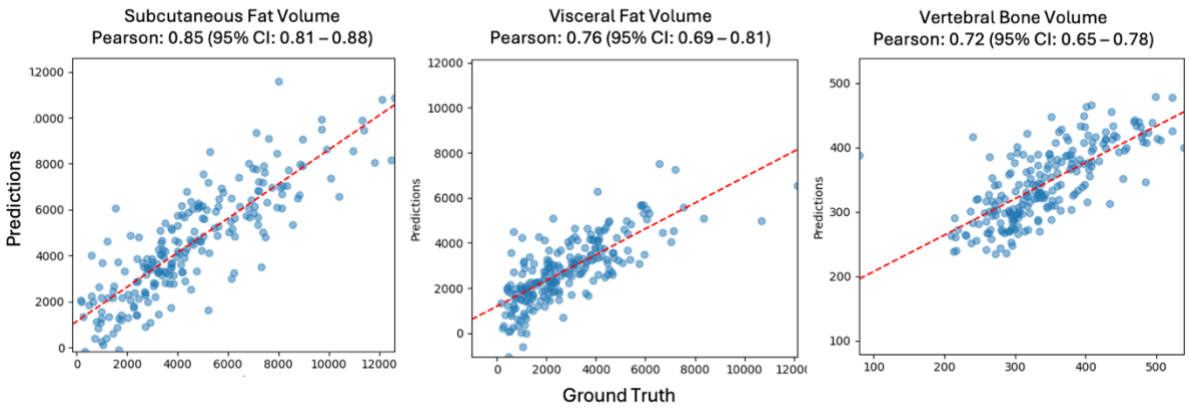

Figure 5.6, Scatter plots showing multimodal model predictions versus ground truth values for three of the top performing body composition metrics.

Table 5.4, Test set performance across various fusion strategies. (Pearson's Correlation)

	Skeletal Muscle Volume	**Visceral Fat Volume**	**Subcutaneous Fat Volume**	**Muscle Radiodensity**	**Fat Free Volume**
Early	0.57	0.69	0.76	0.61	0.5
Intermediate	0.58	0.75	0.82	0.69	0.59
Late	0.58	0.76	0.85	0.68	0.59
	Intramuscular Fat	Aortic Calcification Score (Spearman)	Number of Plaques in Aorta (Spearman)	Skeletal Muscle Index	Visceral Fat Index
Early	0.5	0.61	0.62	0.48	0.7
Intermediate	0.57	0.65	0.67	0.53	0.75
Late	0.55	0.66	0.68	0.53	0.75
	Subcutaneous Fat Index	Fat Free Index	Vertebral Bone Volume	Vertebral Bone Density	
Early	0.8	0.51	0.69	0.34	
Intermediate	0.82	0.6	0.69	0.49	
Late	0.85	0.59	0.72	0.48	

Table 5.5, Test set performance across various fusion strategies. (Pearson's Correlation)

	Skeletal Muscle Volume	**Visceral Fat Volume**	**Subcutaneous Fat Volume**	**Muscle Radiodensity**	**Fat Free Volume**
Early	0.57	0.69	0.76	0.61	0.5
Intermediate	0.58	0.75	0.82	0.69	0.59
Late	0.58	0.76	0.85	0.68	0.59
	Intramuscular Fat	Aortic Calcification Score (Spearman)	Number of Plaques in Aorta (Spearman)	Skeletal Muscle Index	Visceral Fat Index
Early	0.5	0.61	0.62	0.48	0.7
Intermediate	0.57	0.65	0.67	0.53	0.75
Late	0.55	0.66	0.68	0.53	0.75
	Subcutaneous Fat Index	Fat Free Index	Vertebral Bone Volume	Vertebral Bone Density	
Early	0.8	0.51	0.69	0.34	
Intermediate	0.82	0.6	0.69	0.49	
Late	0.85	0.59	0.72	0.48	

5.4.5 L3 Slice Level Models

Our L3 level model achieved similar performance in the top performing categories to the volumetric body composition (T12-L5) models. *Estimates* on the hold out test set achieved a correlation of 0.81 (0.76 – 0.85) for subcutaneous fat area and 0.74 (0.67-0.80) for visceral fat area.

5.4.6 Explainability

The aggregate explainability saliency maps for the prediction of subcutaneous fat volume, visceral fat volume and vertebral bone volume and skeletal muscle volume can be viewed in **Figure 5.7**.

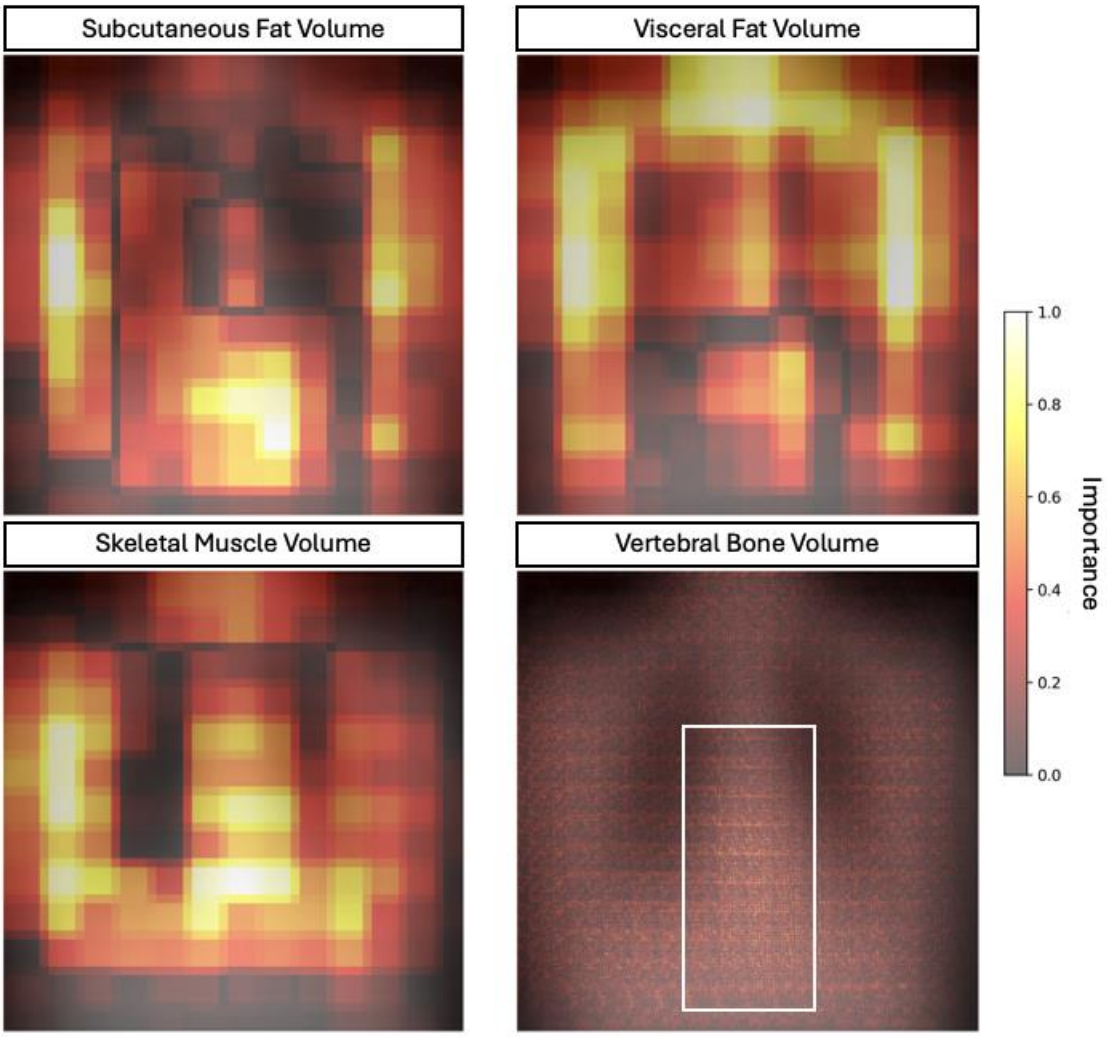

Figure 5.7, Saliency Maps for feature importance in prediction of the top performing body composition metrics. For the first three occlusion-sensitivity saliency map is depicted while for vertebral body volume. The images are aggregates of individual chest radiographs and their respective saliency maps in the hold-out test cohort.

5.4.7 Fairness Analysis

We investigated the performance of our model within different age, sex, and BMI subgroups of our test cohort to identify potential biases in our model. In general, although there was variability in model performance across various subgroups, no definitive trend is observable. The results for the top performing body composition metrics can be seen in supplementary material **Table 5.6**.

Table 5.6, Model performance (AUROC) across different subgroups in the test set. While there is no general pattern, there is variability in model performance with respect to age and BMI.

	Subcutaneous Fat Volume	**Visceral Fat Volume**	**Skeletal Muscle Volume**	**Vertebral Bone Volume**
Age 18 - 39	0.92	0.7	0.73	0.85
Age 40 - 59	0.79	0.71	0.54	0.6
Age 60 - 74	0.81	0.77	0.7	0.75
Age > 74	0.81	0.78	0.48	0.67
Female	0.87	0.75	0.44	0.54
Male	0.82	0.73	0.45	0.50
BMI <25	0.61	0.68	0.52	0.59
BMI 25-30	0.63	0.62	0.57	0.7
BMI 30-35	0.67	0.72	0.68	0.81
BMI >35	0.83	0.68	0.54	0.75

5.5 Discussion

This aim focused on evaluation and comparison of early, intermediate, and late fusion techniques for combining imaging and clinical data on the regression prediction task of estimation of CT-based body composition metrics from chest radiographs. While the previous two aims each had explored multimodal data fusion using different data modalities, here, I attempted address my overarching question of evaluation of multimodal fusion approaches in predictive medical models by exploring the fusion of 2D imaging and clinical variables to develop a deep learning model designed to predict continuous variables relating to body composition. In addition, given the poor performance observed from early fusion in the last aim, I explored using another approach in early fusion in which instead of extracting features from the imaging using pretrained networks, I expanded the dimensionality of clinical variables by generating embeddings from them which I added to the images. In this aim, I demonstrated that late fusion outperformed other fusion strategies closely trailed by intermediate fusion. This aim further confirms our findings that early fusion does not perform well when combining imaging and clinical variables.

In this study we developed a multimodal deep learning model using chest radiographs and four clinical variables to estimate various body composition measures. Our model achieved good predictive performance in estimating subcutaneous adipose tissue, visceral adipose tissue, vertebral bone volume and to a lesser extent skeletal muscle volume. The resulting model can be used for a larger population of people to get estimated body composition metrics that are more accurate surrogates compared to weight and BMI alone while don't involve costly imaging or large amounts of radiation.

Our multimodal model outperformed unimodal clinical-only and imaging-only models with performance increase being especially noticeable on estimating subcutaneous fat volume visceral fat volume and vertebral bone volume. The combined model did not perform noticeably well on

tissue level body composition metrics or aortic calcification score levels. We also trained models to predict mid-L3 level body composition metrics. We showed that the models for mid-L3 body composition metrics perform similar to the volumetric ones.

Our clinical only model that included sex at birth, weight, height, and age could be used to estimate some body composition measures as well, especially subcutaneous fat area, skeletal muscle area and aortic calcification score. However, a clinical only model performs poorly when estimating other body composition metrics, especially visceral fat area. Visceral fat area is particularly important as it is correlated with adverse outcomes in various diseases (83,85).

The ability to differentiate subcutaneous and visceral fat is specifically important due to the importance of visceral fat volume in prediction of various risk factors from cardiometabolic disease to cancer (83,86). Another important body composition metric which is proven to be a risk factor for adverse outcomes is sarcopenia (99). We looked at both skeletal muscle volume, radiodensity and intramuscular fat in our study. While our final model only achieved moderate correlation in predicting skeletal muscle volume, during training and validation we observed more promising performance. This decline in performance could be attributed to overfitting of the model or could be due to other differences between the training and testing cohorts. In addition, our model had moderate to good performance in estimating the muscle radiodensity ($r = 0.69$, $0.61 – 0.75$). Additional work is required to explore the possibility of estimating these two metrics from radiographs.

5.5.1 Importance of fusion timing in final model performance

Previous studies that investigated the timing of fusion for medical data, have found conflicting answers. For example, study that looked into developing a multimodal model for the detection of

pulmonary embolism concluded that the model that used late fusion performed best (96). Other studies, looking at fusion of various types of EHR data or imaging have found that intermediate fusion performed best for their study (24,80).

In our study, we found that while the performance of late and intermediate based fusion models was very close, the late fusion model had a higher performance. Interestingly, for the L3 level body composition metrics, the performance of the intermediate fusion model was higher. These differences show the importance of experimenting with different fusion strategies in multimodal deep learning studies. Various factors might be influencing the model performance across different fusion strategies. For example, in late fusion the imaging only and the clinical only models are trained and then their weights are frozen. This will lead to a simpler model at the end that is less likely to overfit to the training data due to the lower number of trainable parameters. This approach is specifically useful in cases in which the training sample size is small. On the other hand, early and intermediate fusion strategies allow the network to learn patterns across the various modalities.

5.5.2 Model explainability, fairness and feature importance

The saliency maps in **Figure 5.7** show that for the estimation of subcutaneous and visceral fat tissue volume, the model looks at the mediastinum (which contains visceral adipose tissue), the suprasternal region and the sides of the trunk. Although the saliency maps overlap, they put different importance on different parts. For example. The visceral fat volume model is more focused on the mediastinum and neck region compared to the subcutaneous fat volume model. The model for vertebral bone volume focuses on the part of the image that contains the lower thoracic vertebrae behind the mediastinum. In general, our saliency maps show that the model is looking into the appropriate regions of the radiograph when estimating body composition metrics.

Our fairness analysis showed that our model had variable performance across different age and sex groups. While these differences did not point to a definitive trend, it is important to note them when applying this model to other populations.

5.5.3 Failure Analysis

Three cases that were among the worst performing in the test cohort can be seen in **Figure 5.8**. Compared to a typical chest radiograph, these demonstrate problems like incomplete coverage of the trunk on both sides, patient rotation, inadequate lung inflation and low exposure. Based on these findings, we hypothesize that our model's performance could be higher if provided with high quality chest radiographs.

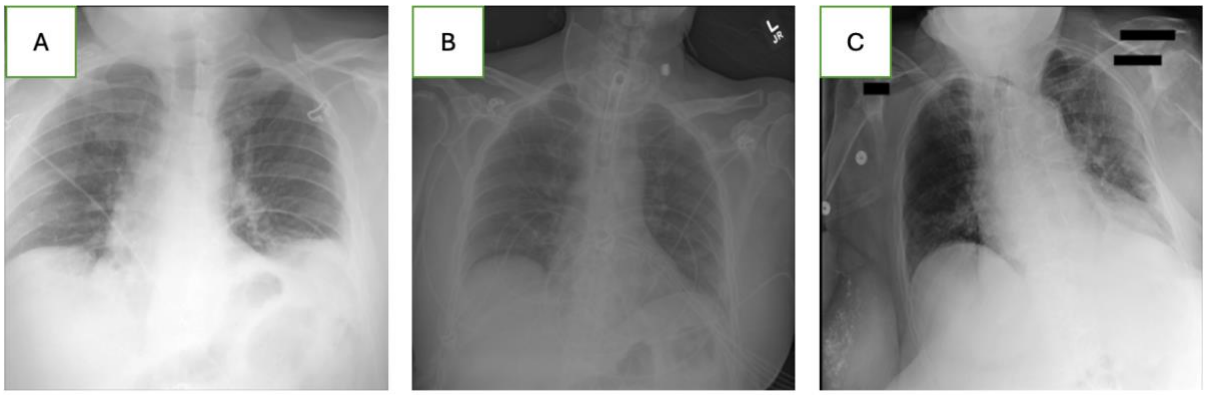

Figure 5.8, Three samples of the of the radiographs in which the model had poor performance. A. Parts of the chest radiograph that are important to the model are missing from the image. B. This chest radiograph is dark. C. The patient is rotated, and the lungs are not properly inflated.

5.5.4 Challenges with Multimodal Fusion

Our findings in this chapter are in concert with findings of the previous two chapters and previous studies that when compared, intermediate and late fusion strategies often outperform early fusion

(24,96). However, I faced challenges in implementation of each of these fusion strategies that required specific interventions to tackle. The choice of these strategies may have implications on the model performance.

For early fusion, in cases where the initial data types are not readily compatible (e.g. imaging and structured data) steps are needed to bring these data types into the same space for them to be analyzed together. Commonly a feature extractor (pretrained convolutional neural networks or radiomic feature extractors for imaging data, pretrained language models for natural text, polygenic risk score for genomic data, etc.) is used to convert these data types into variables that can be concatenated or added together. Preferably little to no learning must happen at this stage for the fusion strategy to be considered 'early fusion'. This hinges the ability of early fusion model on the original feature extractor's ability to extract meaningful features for problem at hand. As most feature extractors are usually general-purpose feature extractors and are not specifically trained for the problem of interest, the features they generate may not be the optimum set of features necessary for our prediction task. It is possible to tackle this problem using question specific feature extractors trained on outside data. Another possible solution is to try to bring the modality with a smaller number of dimensions (here, the 4 clinical variables), features to become compatible with the larger modality (here this would be the imaging). In this chapter, we used this strategy to use a small fully connected layer to generate an embedding from the four clinical variables that were subsequently added to the imaging before further processing.

For intermediate fusion, the main challenge I faced was that the heterogeneity of the network and its large size (intermediate fusion has the largest number of parameters being simultaneously trained together) could cause problems like overfitting and unstable training. I tried to tackle these issues by introducing more strict regularization and comprehensive hyperparameter tuning.

Furthermore, in intermediate fusion, the multiple parts of the network that analyze different data modalities may require different hyperparameters to train effectively. Using different hyperparameters (e.g., learning rate, weight decay, dropout rate) for the various parts of the network is not currently readily possible and would also complicate hyperparameter tuning. To address this challenge, like the previous chapters, I resorted pretraining each network component in a unimodal manner and then loaded the weights into the intermediate fusion network for further finetuning.

5.5.5 Limitations

Although we achieved good performance on multiple tasks in this study, due to computational constraints and data loss, the sample size of our study was relatively small for this task. Additionally, the maximum 90 days difference between the chest radiograph and CT scan could potentially introduce noise into our data *because* body composition could change rapidly due to various factors like diet, exercise, or disease. Another limitation of our study is using abdominal CT scan-based body composition metrics as opposed to whole-body scans. Finally, poor performance of the model on estimating Fat Free Index could *be* attributed *in some part* to the random difference between train and test sets as was determined in our ANOVA analysis.

5.6 Conclusion

In this aim, we furthered our understanding of multimodal data fusion in predictive clinical models by comparing the implications of different fusion strategies when combining 2D imaging and clinical data for a deep learning regression task. In conclusion, we developed a multitask multimodal deep learning model to estimate various body composition metrics, including subcutaneous fat volume, visceral fat volume and vertebral bone volume using a chest radiograph

and four relevant clinical variables. In this study, we demonstrated the incremental value of combining clinical and imaging information in prediction of body composition metrics. This is in alignment with our previous findings in 4 that adding various modalities improves performance beyond what is achievable when single modalities are used. Furthermore, we demonstrated that late fusion achieved superior performance in prediction of volumetric body composition metrics. Intermediate fusion followed closely with the difference in performance not statistically significant in many cases. However, early fusion was less effective in this prediction task. These findings are in concert with our findings in the previous chapters in which intermediate and late fusion were again the top performing models. Over the next chapter, I will summarize my findings across this body of work on evaluation of fusion approaches in predictive medical models. A summary of lessons learned about fusion in this aim can be seen in **Figure 5.9**.

Early Fusion	Intermediate Fusion	Late Fusion
• Good choice if features can be represented with no simplification/dimensionality reduction. • Can handle missing data if appropriate architecture is used. • May lose some signal when feature extractors are used. • May help when sample size is very small for finetuning on imaging. • Performed worse even when structured data was embedded in imaging.	• May overfit to training data due to large number of trainable parameters. • Better for combination of imaging and clinical data. • Various network parts train at different speeds and may need different hyperparameters. Pretraining helps with stable training. • More design choices compared to other methods. • More costly and resource intensive.	• Can combine diverse networks like XGBoost and transformer. • Does not learn inter-modal correlations. • Less overfitting due to lower number of trainable parameters. • Performance depends on unimodal models. • Better performance when all modalities have good performance.

Figure 5.9 Lessons learned from Aims 1,2 and 3.

6. Summary of Aims

6.1 Introduction

Deep learning medical predictive models can improve care by providing patient-tailored decision support and uncover insights about diseases. Traditionally, deep learning models often used data from a single modality to make predictions about a patient (i.e. deep learning model to diagnose pneumonia using a chest radiograph). However, the use of unimodal data may hamper the model's ability to make accurate predictions as often data from different modalities are required to interpret a patient's condition (i.e. radiographic signs of pneumonia in addition to fever and respiratory symptoms). We can observe an increasing trend in the use of multimodal deep learning predictive models in medicine due to the benefits that come with multimodal modeling. Multimodal data enables the model to learn both marginal (intramodal) and intermodal relationships, gaining a more comprehensive understanding of the question at hand. The use of multimodal data in predictive modeling is closer to actual human practice in which an expert will use data from multiple modalities (i.e. imaging, clinical history, laboratory test results, …) to make decisions about a patient or understand underlying mechanisms of disease. Categories of medical data include structured and unstructured electronic health records data, imaging data, multi-omics, survey data, wearables, and population level data.

Effective development of medical multimodal deep learning models is dependent upon addressing challenges inherent to multimodal modeling in medicine. These include addressing the dimensionality and size differences between different modalities that can range from only a few clinical variables to hundreds of gigabytes for an individual's whole genome sequence data; data or modality missingness, and diverse analytic methods.

The process of integration of multiple data modalities in deep learning modeling is called *data fusion*. Three main approaches to data fusion have been introduced. These include early or feature level fusion, intermediate fusion and late or decision level fusion. Early fusion involves the concatenation of raw input features from different modalities or features/variables extracted from them. In intermediate fusion, modality specific network sections extract high-level features as form of matrix embeddings from each modality. These embeddings are then added together using methods like concatenation or cross-modal attention. Later network sections analyze these features together resulting in learning of inter-modal relationships even in modality specific layers. Late fusion strategy involves training of modality specific models the predictions of which are used to make the final prediction. This can be done via voting mechanisms or another smaller neural network that takes as input the predictions for each modality.

Few studies have explored the implications of each of these fusion strategies in medical deep learning models. Studies in other domains have shown that early fusion networks are good at extracting fine-grain relationships between input variables from different modalities while intermediate fusion networks are better at finding inter-modal relationships between more complex features. In medicine, studies that have explored different fusion approaches have come to varying conclusions regarding the best performing fusion strategy.

Through the course of this dissertation, I plan to address the current gap in literature by exploring the impact of choice of fusion strategy in multiple medically relevant deep learning projects. In my first aim titled ‘Evaluation and comparison of early, intermediate, and late fusion techniques for combining exposures, clinical and genomics data for disease risk prediction task using All of Us: Risk of CKD in patients with type 2 diabetes’ I explore the gap in the literature regarding fusion of genomic, longitudinal EHR and survey data specifically in the context of using

transformer models. In the second aim titled ‘Development and assessment of the incremental value of combining a deep convolutional neural network feature extractor on imaging data and clinical data on a binary prediction task: Predict post-surgical margin status in soft tissue sarcoma’ in further explore fusion by looking at fusion of longitudinal cross-sectional imaging data with clinical data extracted from the EHR. This aim is focused on the gap in literature regarding fusion of imaging and clinical data. Finally, while the previous aims and most previous studies in literature focus on fusion in problems with binary outcomes, in Aim 3 titled ‘Evaluation and comparison of early, intermediate, and late fusion techniques for combining imaging and clinical data on a regression prediction task: Estimation of CT-based body composition metrics from chest radiographs’, I explore fusion of 2D imaging and clinical variables in the context of a multitask deep learning model for a regression task to estimate body composition metrics.

6.2 Aim 1, Evaluation and comparison of early, intermediate, and late fusion techniques for combining exposures, clinical and genomics data for disease risk prediction task using All of Us: Risk of CKD in patients with type 2 diabetes.

In this aim, to address the gap in literature regarding the use of fusion strategies to combine genetic, EHR and survey data. Here, I used the NIH’s All of Us dataset to predict risk of progression to chronic kidney disease (CKD) in patients with type 2 diabetes (T2D). Progression to CKD is one of the most detrimental adverse effects of T2D. Progression to CKD can be slowed using preventive strategies including tight glycemic control. However, these strategies come with other risks. Individualized approaches to treatment of each patient based on their medical history and risk factors is warranted to minimize risks and maximize benefits. While previous studies have focused on using unimodal data to predict risk of progression to CKD in T2D patients, the integration of genomic and survey data into these models has been underexplored. In addition,

there is a gap in literature regarding the utility of early, intermediate, and late fusion approaches in combining genomic, longitudinal EHR and survey data in deep learning models in general. The EHR data constituted a sequence of events recorded in the EHR in addition to the time difference between each consecutive record and the value assigned to that record (e.g. result of the laboratory test). The genomic data consisted of two polygenic risk scores from the literature for CKD and the list of pathogenic and likely pathogenic variants for CKD extracted from ClinVar. Finally, the survey data consisted of responses to the family history section of the personal and family medical history survey, SDOH survey and the lifestyle survey.

For early fusion, only positive responses from the survey, in addition to the two polygenic risk scores and positive SNPs were added to the beginning of the sequence for EHR data. These were passed to the time and value aware transformer model for analysis. For intermediate fusion, two shallow neural networks were used to generate embeddings from the genomic and survey data. These were added to the last hidden layer of the transformer model and passed through a series of fully connected layers at the end. For late fusion, XGBoost was used to make models based on genomic and survey data and the transformer model was used to make a model based on EHR data. The predictions of all these networks were concatenated and passed through a fully connected layer for the final prediction.

6.2.1 Key Contributions

We demonstrated that the time and value aware transformer was able to identify patients at a higher risk for progression to CKD within the next 5 year with an AUROC of 0.73. Additionally, the survey only model achieved an AUROC of 0.59 while the genomic model achieved an AUROC of 0.53.

In this aim, we introduced time and value-aware transformer based fusion methods that can be used to integrate genomic, EHR and survey data using early, intermediate, and late fusion strategies. To the best of my knowledge, this is the first work combining all these elements together. The performance of the multimodal models after the addition of genomic and survey data did not significantly improve, highlighting the limitations of multimodal modeling when some modalities have low predictive power. The lack of improvement of performance hinders our ability to assess which fusion strategy performed best in this setting. Nevertheless, the early fusion network achieved slight performance advantage over the validation set while being less overfit to the training set.

We demonstrated that the use of early fusion with the transformer model was feasible due to the ability to only include none-zero variables. This enabled the early fusion network to use the input variables without any extra loss of signal attributed to feature extraction strategies to enable adding different modalities together. The intermediate fusion network had the lowest overall performance and highest level of overfitting to the training data, highlighting the complexities of developing intermediate fusion networks and their propensity to overfit to the training data due to the higher number of trainable parameters.

6.2.2 Limitations

We faced multiple limitations in this aim. First, there were multiple design choices for the time and value aware transformer model that could be tuned to optimize performance. However, due to computational and cost constraints for this project, I could not explore all the possible options. These included but were not limited to the base transformer model architecture, the maximum sequence length, padding and truncating strategy, time and value embedding generation strategy and vocabulary tokenization strategy. Additionally, I was not able to perform a comprehensive

hyperparameter tuning over the entire hyperparameter space due to computational constraints. Finally, each fusion approach has variants that may inadvertently influence its performance. For example, in intermediate fusion, the choice on the level after which fusion occurs can influence model performance. In late fusion, the choice of ensembling the model predictions may influence final performance.

6.2.3 Future Directions

These design choices, while outside the scope of this work, offer promising future directions to advance both the use of transformer models in analyzing biomedical data and fusion of different data modalities. Finally, given the low influence of survey and genomic data in this question, a promising approach would be to apply the same strategy in another disease in which the role of genetics and survey data is more prominent.

In summary, in Aim 1, we introduced a novel multimodal time and value aware transformer architecture and evaluated the impact of fusion timing in this network. In terms of my overarching question, in this aim, I explored fusion of clinical, genomic and survey data in the context of a binary prediction task. However, given the diversity that exists in biomedical data, it is imperative to study fusion when other modalities are involved as a next step. Hence, in the next aim, we explored fusion of longitudinal cross-sectional imaging data with clinical variables extracted from the EHR for another binary prediction task.

6.3 Aim 2, Development and assessment of the incremental value of combining a deep convolutional neural network feature extractor on imaging data and clinical data on a binary prediction task: Predict post-surgical margin status in soft tissue sarcoma.

To answer the overarching question of evaluating the multimodal data fusion approaches in medical predictive models, I extend the concepts introduced in Aim 1 by adding new data modalities and a new medically relevant problem. In this aim, I used fusion approaches with longitudinal and cross-sectional medical imaging data, in conjunction with clinical and pathology variables extracted from the EHR and evaluation of post-excision tumor samples by pathologists to predict involvement of post-surgical margin in patients with soft tissue sarcoma. Soft tissue sarcomas are a rare and diverse type of cancer originating from soft tissues including connective tissue, muscles, and adipose tissue. Their treatment often involves neoadjuvant radiotherapy often in addition to chemotherapy or immunotherapy and subsequent surgical excision of the tumor. One of the most important predictors of recurrence in soft tissue sarcomas is the involvement of post-surgical margins meaning that viable tumor cells were left in the body despite of the excision of the tumor. There exist no established risk scores to identify patients at a high risk for involvement of post-surgical margins. In this aim, we developed a multimodal deep learning model using early, intermediate, and late fusion approaches to estimate this risk using longitudinal cross-sectional MR imaging of the tumor before excision and select clinical and pathology variables.

Early fusion was implemented by using a pretrained 3D ResNet10 model to extract imaging features from the pre- and post-neoadjuvant therapy MRIs of the tumor region. These features were concatenated with the clinical and pathology variables and passed through a neural network to make the final prediction. For intermediate fusion, the imaging part of the network consisted of a 3D ResNet10 model the final embedding layer of which was concatenated with the embeddings generated using a shallow neural network from the clinical data. These embeddings were passed through the rest of the network to make the final prediction. For the late fusion strategy, two

separate imaging and medical models were trained using the same CNN network architecture and a shallow neural network respectively.

We demonstrated that the addition of clinical and imaging data resulted in multimodal models that achieved higher performance compared to the unimodal models. Our findings demonstrated that the intermediate fusion network achieved the highest performance in cross-validation and test sets compared to early and late fusion networks.

6.3.1 Key Contributions

In this aim, we experimented with fusion strategies used in literature to integrate imaging and structure clinical variables and demonstrated that in this scenario, intermediate fusion outperformed early or late fusion approaches. This is interesting as among the 3 aims of this dissertation, this aim involved the least number of cases (202 patients). This highlights the importance of types of interactions between the different modalities in selection of the proper fusion strategies. In this scenario, variables like the location and histological subtype of the tumor are associated with the imaging characteristics of the tumor and the intermediate fusion approach enables the model to more easily associate the clinical variables with these complex imaging features as described in previous literature.

Additionally, we demonstrated how the multimodal model outperforms a model based on the same clinical variables and imaging features manually extracted by expert radiologists.

Apart from the objective findings that intermediate fusion outperformed other fusion strategies, there were implementations challenges that shed light on other aspects of multimodal data fusion that I faced in this aim. First, the intermediate fusion network was more challenging to train and initially often overfit during cross-validation. This behavior is expected due to the low sample size

and higher number of trainable parameters. In fact, I realized that in intermediate fusion, different network sections may train optimally using different hyperparameters (i.e. learning rate, weight decay,…). To overcome this challenge, I used unimodal pretrained weights in each network section instead of other methods of initialization of the network and then let the network to fine-tune using lower learning rates.

6.3.2 Key Limitations

One limitation of this aim was that longitudinal images were not available for all the cases. Instead of removing those cases from analysis, we included them during training by passing blank images in cases where imaging data was not available. However, we separately reported the model performance on the longitudinal section of the test set.

Another limitation of this aim was that like Aim 1, early, intermediate, and late fusion approaches can be further customized. For example, one could experiment with different feature extractors for the imaging data (e.g. radiomics, other pretrained networks, …). In addition, similar to Aim 1, other variations could be used for intermediate and late fusion as well.

6.3.3 Future Directions

Given the findings of this aim, future directions could include further experimenting with strategies to optimize intermediate fusion with respect to timing of the fusion and strategy used to fuse different modalities. Of particular interest is use of cross-modal attention mechanisms instead of simple concatenation of embeddings as it will allow the model to focus on specific features depending on data from other modalities and has shown promising performance in other domains (24). Another avenue of research would be exploring the level at which the fusion of modalities happens. In this work, I concatenated the embeddings from different modalities from the final layer

of each of the modalities. Additionally, one could experiment with different feature extractors for early fusion including experimenting with other pre-trained networks or rule-based strategies like radiomics.

Finally, given the importance of margin detection in soft tissue sarcoma, future steps in this project could include external validation of the models and prospective studies to establish the risk model's ability to prevent positive margins in patients with soft tissue sarcoma.

In summary, in Aim 2, we explored the fusion of cross-sectional imaging data with clinical and pathology variables in deep learning models to predict a binary outcome. We demonstrated that intermediate fusion was the best performing model. While in aims 1 and 2, we explored fusion of multiple data modalities including EHR, pathology, genomics, survey results and imaging, both aims were focused on binary outcome prediction tasks. To further explore our overarching question of evaluating data fusion approaches in predictive medical models, and to investigate the role of different outcome types, in Aim 3, we will look into fusion of 2D imaging and clinical variables in the context of a regression problem to estimate multiple body composition metrics using chest radiographs.

6.4 Aim 3, Evaluation and comparison of early, intermediate, and late fusion techniques for combining imaging and clinical data on a regression prediction task: Estimation of CT-based body composition metrics from chest radiographs.

As described in the last section, this aim focuses on answering the overarching question of evaluating multimodal data fusion approaches in medical prediction models when fusion is used for a regression task to estimate multiple continuous variables. To answer this question, in this aim we explore fusion of 2D chest radiograph images with four relevant clinical variables to estimate

CT-based body composition metrics. Body composition metrics are quantitative measures of various body components and can include factors like visceral and subcutaneous fat volume, skeletal muscle volume, and skeletal muscle density. Studies have shown that body composition metrics are implicated in prognosis of various conditions. For example, visceral adipose tissue volume is associated with risk of cardiovascular disease and diabetes. In addition, lower skeletal muscle volume is associate with worse prognosis in some cancers like lung cancer. Current methods to calculate body composition metrics include performing MRIs, CT scans, Dual X-Ray Absorptiometry (DXA) and electrical impedance analysis. These methods suffer from high costs, need for specialized equipment and potential for exposure to excessive radiation. Hence, it would be beneficial if body composition metrics could be estimated from commonly available medical data. In this aim, I evaluated data fusion strategies to develop a multitask deep learning model to estimate various body composition metrics (continuous variables) using a combination of a chest radiograph and four clinical variables including age, sex, height, and weight.

For early fusion, in this aim I explore another variation in which I add embeddings generated from the clinical variables using a fully connected layer into the image before passing through the rest of the network which is a ResNet10 convolutional neural network. This way, I avoid the need to use feature extraction techniques which may undermine the early fusion methods ability to learn task specific imaging features.

The implementation of intermediate and late fusion networks in this aim was similar to Aim 2 by using a shallow neural network to generate embeddings or predictions from the clinical data and a ResNet10 network to generate embeddings/predictions from the imaging data. However, in this aim, the neural network was trained using the Huber loss, which is a mixture of mean squared error loss and mean absolute error loss. These models estimated multiple body composition metrics as

continuous variables and the performance was measured by calculating the correlation between the ground truth values and predicted values.

6.4.1 Key Contributions

In this aim, we demonstrated that late fusion outperformed other fusion methods in estimation of almost all body composition metrics. This was closely followed by intermediate fusion. However, similar to Aim 2 and in contrast to Aim 1, early fusion did not perform well on this task. The better performance of the late fusion method could be explained by the fact that both clinical and imaging only models had similarly good performance in estimation of body composition metrics. This is in contrast with Aim 2 in which the performance of the imaging model was much higher compared to the clinical model. In addition, there could be more cross-modal interaction between the modalities in Aim 2 compared to Aim 3, which resulted in poorer performance of late fusion in that aim.

The fact that late fusion achieved consistently higher performance in estimation of multiple body composition metrics shows that the choice of fusion may depend more on the characteristics of the input data rather than the characteristics of the outcome. Otherwise, it would be reasonable to expect that the best performing fusion strategy be different for different body composition metrics.

Similar to Aim 2, to avoid overfitting and poor training performance in the intermediate fusion model, we had to train individual network sections in a unimodal fashion and then fine tune the resulting weights. This behavior can be attributed to the observation that different network parts in intermediate fusion train at different speeds and with different hyperparameters.

Finally, this Aim resulted in a publicly available tool for estimation of body composition metrics using chest radiographs and four relevant clinical variables.

6.4.2 Key Limitations

A major limitation of Aim 3 was the low number of training samples which may hinder the model's generalizability. This was partly due to problems with retrieving the images or clinical data, or calculation of ground truth body composition metrics from CT scans. Given the retrospective nature of this study, this level of loss of samples due to these reasons is not unexpected, especially since these scans were performed for clinical indications and not for research. We utilized data augmentation techniques to address this limitation to some extent.

Another limitation of this study was the use of retrospective data, which was not necessarily collected for body composition calculation. Hence, some body composition metrics, notably those that relied on CT Hounsfield unit (HU) values, may have noise that resulted in poor performance for those metrics (e.g. skeletal muscle radiodensity, vertebral bone density). However, this should not have affected volumetric metrics like subcutaneous and visceral fat volume.

6.4.3 Future Directions

Similar to aim 2, future directions could include further exploration of different variations of intermediate and late fusion techniques. Notably, while we used a small neural network to aggregate the two unimodal models in late fusion, other simpler strategies like voting techniques may also result in good performance. In intermediate fusion, we could further explore the level at which the fusion happens. Finally, exploring other deep learning methods, including the more recent vision-transformers could show promising results.

With respect to the body composition estimation model, next steps can include external validation of the model in data extracted from outside institutions. Additionally, studies can investigate the benefit of using these estimated body composition metrics compared to commonly used measures

of obesity like weight and body mass index (BMI). Finally, the use of quantitative CT scans can provide more accurate ground truth values.

In summary, in aim 3, we evaluated the implications of using early, intermediate and late fusion techniques in a regression prediction task. We demonstrated that intermediate and late fusion strategies outperform early fusion with the late fusion strategy achieving slightly higher performance in this case. Our findings were in agreement with findings from the previous aims that in combining imaging and clinical data, early fusion may not be able to extract cross-modal associations effectively.

6.5 Key Contributions Across the Aims

In this body of work, I explored and evaluated the performance and implications of data fusion strategies in developing predictive medical models across various medical data modalities and relevant medical problems. In Aim 1, e*valuation and comparison of early, intermediate, and late fusion techniques for combining exposures, clinical and genomics data for disease risk prediction task using All of Us: Risk of CKD in patients with type 2 diabetes,* I explored fusion of genomic, EHR and survey responses data using a transformer-based model. In that aim, I introduced multimodal transformer-based architectures based on early, intermediate, and late fusion strategies to combine these data types. While the results of this aim were inconclusive regarding which fusion strategy performs best in that scenario, early fusion had a slight advantage. In Aim 2, *development and assessment of the incremental value of combining a deep convolutional neural network feature extractor on imaging data and clinical data on a binary prediction task: Predict post-surgical margin status in soft tissue sarcoma,* I extended the concepts studied in Aim 1 by using fusion strategies to combine imaging, clinical and pathology variables. In that aim, I demonstrated how intermediate fusion outperforms other methods in prediction of post-surgical margin status closely

followed by late fusion. Finally, in Aim 3, *evaluation and comparison of early, intermediate, and late fusion techniques for combining imaging and clinical data on a regression prediction task: Estimation of CT-based body composition metrics from chest radiographs*, I explored fusion in the context of regression task to estimate multiple body composition metrics while using imaging and clinical variables. In that aim, I again demonstrated that intermediate and late fusion approaches can achieve superior performance compared to early fusion when it comes to fusion of imaging and clinical data. A summary of the characteristics of each aim can be seen in **Figure 6.1**.

Aim	Data Types	Sample Size	Outcome Type	Network Types	Outstanding features	Best Fusion Strategy
Aim 1: Prediction of Progression to CKD	Longitudinal EHR Whole genome Sequence Survey	~40,000	Binary	Transformers Shallow Neural Networks XGBoost	Genomics and survey had small predictive value All data could be integrated into early fusion without loss of information.	Not able to assess (Maybe early due to less overfitting.)
Aim 2: Prediction of Post-surgical Margin Involvement in STS	Multi-timepoint Cross-Sectional Imaging Structured Clinical Pathologic	~200	Binary	3D Convolutional Neural Networks. Shallow Neural Networks.	Imaging data was more complex. Imaging had more contribution to prediction. Sample size was small.	Intermediate
Aim 3: Estimation of Body Composition	2D Imaging Structured Clinical	~1000	Continuous	2D Convolutional Neural Networks. Shallow Neural Networks.	Both modalities had high contribution to the outcome. Sample size was average.	Late closely followed by Intermediate

Figure 6.1 The data and question characteristics for each of the aims of the dissertation with respect to factors that may be important in the choice of fusion.

In summary, insights from my work show that early fusion can be beneficial in scenarios in which the raw data can be used without the need for additional feature extraction or dimensionality reduction techniques. This was possible in Aim 1, in which I could add positive survey responses and SNPs to the sequence passed through the transformer model, without losing any data. However, in scenarios in which I had to use a feature extraction technique (like in Aim 2) to extract

features from the imaging data, the model performance suffered. Likely due to the fact that those features were not specific to the task at hand. Early fusion also performed poorly in Aim 3, in which I attempted to address this issue by adding embeddings from the clinical data to the raw imaging data. This was expected as previous literature on early fusion on other domains had demonstrated that early fusion of raw values is not as effective in finding cross-modal interactions, which are often identified in later layers of the network.

Apart from performance metrics, we identified multiple implications for each of the fusion strategies. Early fusion strategy is the most frequently used method in literature due to its simplicity. Early fusion networks often have a smaller number of trainable parameters and are less likely to overfit to the data as was the case in all our aims. This attribute can be beneficial in scenarios in which there exists sample size issues or compute constraints. The late fusion approach also benefits from a lower number of trainable parameters compared to intermediate fusion. However, with late fusion, unimodal networks need to be pretrained, so the overall number of trainable parameters is often higher than with early fusion (in which usually a feature extraction strategy is used to simplify the input data). One advantage of late fusion, demonstrated in Aim 1, is the ability to integrate different types of models. This may be beneficial in scenarios in which a certain type of modeling approach that works very well for one modality is not compatible with other parts of the network (e.g. XGBoost models and transformer models).

Finally, I observed that the intermediate fusion networks are the hardest to train due to multiple factors. First, they have the highest number of simultaneously trainable parameters, which predisposes them to overfitting. Second, the various network parts for the different modalities may need different hyperparameters to train effectively. This is normally not straightforward to

implement in practice and the method that I used to get around that limitation was to use unimodally pretrained weights in the intermediate fusion network.

In conclusion, the choice of fusion can have significant influence on the overall performance of deep learning models in medicine. Some factors that may be implicated in making a decision about the fusion strategy are the sample size, the differences in the dimensionality of the input data and the need for feature extraction methods, the unimodal performance of each of the modalities and the level of cross-modal interaction between the variables. In general, intermediate, and late fusion approaches are recommended for combining imaging with clinical data. Intermediate fusion may be preferred in cases in which there is a higher expectation for cross-modal interaction between the clinical variables and the complex imaging features while late fusion may be preferred when each modality already achieves high predictive performance. Early fusion could be beneficial in cases with low sample size, or scenarios in which additional feature extraction methods are not required to bring the data from different modalities into the same dimensions. A summary of the findings can be viewed in **Figure 6.2**.

Early Fusion	Intermediate Fusion	Late Fusion
• Good choice if features can be represented with no simplification/dimensionality reduction. • Can handle missing data if appropriate architecture is used. • May lose some signal when feature extractors are used. • May help when sample size is very small for finetuning on imaging. • Performed worse even when structured data was embedded in imaging.	• May overfit to training data due to large number of trainable parameters. • Better for combination of imaging and clinical data. • Various network parts train at different speeds and may need different hyperparameters. Pretraining helps with stable training. • More design choices compared to other methods. • More costly and resource intensive.	• Can combine diverse networks like XGBoost and transformer. • Does not learn inter-modal correlations. • Less overfitting due to lower number of trainable parameters. • Performance depends on unimodal models. • Better performance when all modalities have good performance.

Figure 6.2 Lessons learned about fusion across the three aims of this dissertation.

6.6 Future Directions

My findings resulted in multiple hypothesis regarding the utility of different types of fusion in predictive medical models. A future direction could include a more systematic review of these factors in a large synthetic dataset in which the various characteristics of the data can be altered and experimented with. These could include evaluating fusion across varying sample sizes, using different combinations of input data modalities and various levels of interaction between raw variables from each modality or complex features extracted from modalities.

7. References

1. The “All of Us” Research Program. N Engl J Med. 2019 Aug 15;381(7):668–76.

2. Multimodal biomedical AI | Nature Medicine [Internet]. [cited 2025 Jun 10]. Available from: https://www.nature.com/articles/s41591-022-01981-2?utm_source=chatgpt.com

3. Comparison of deep learning with traditional models to predict preventable acute care use and spending among heart failure patients | Scientific Reports [Internet]. [cited 2025 Jun 10]. Available from: https://www.nature.com/articles/s41598-020-80856-3?utm_source=chatgpt.com

4. Kline A, Wang H, Li Y, Dennis S, Hutch M, Xu Z, et al. Multimodal machine learning in precision health: A scoping review. Npj Digit Med. 2022 Nov 7;5(1):171.

5. Person M, Jensen M, Smith AO, Gutierrez H. Multimodal Fusion Object Detection System for Autonomous Vehicles. J Dyn Syst Meas Control [Internet]. 2019 May 8 [cited 2025 May 17];141(071017). Available from: https://doi.org/10.1115/1.4043222

6. El-Ateif S, Idri A. Multimodality Fusion Strategies in Eye Disease Diagnosis. J Imaging Inform Med. 2024 Apr 19;37(5):2524–58.

7. Roest C, Yakar D, Rener Sitar DI, Bosma JS, Rouw DB, Fransen SJ, et al. Multimodal AI Combining Clinical and Imaging Inputs Improves Prostate Cancer Detection. Invest Radiol. 2024 Dec;59(12):854.

8. Stahlschmidt SR, Ulfenborg B, Synnergren J. Multimodal deep learning for biomedical data fusion: a review. Brief Bioinform. 2022 Mar 1;23(2):bbab569.

9. Boehm KM, Khosravi P, Vanguri R, Gao J, Shah SP. Harnessing multimodal data integration to advance precision oncology. Nat Rev Cancer. 2022 Feb;22(2):114–26.

10. Huang SC, Pareek A, Seyyedi S, Banerjee I, Lungren MP. Fusion of medical imaging and electronic health records using deep learning: a systematic review and implementation guidelines. Npj Digit Med. 2020 Oct 16;3(1):136.

11. Kohane IS, Aronow BJ, Avillach P, Beaulieu-Jones BK, Bellazzi R, Bradford RL, et al. What Every Reader Should Know About Studies Using Electronic Health Record Data but May Be Afraid to Ask. J Med Internet Res. 2021 Mar 2;23(3):e22219.

12. Qoku A, Katsaouni N, Flinner N, Buettner F, Schulz MH. Multimodal analysis methods in predictive biomedicine. Comput Struct Biotechnol J. 2023 Nov 20;21:5829–38.

13. Shickel B, Tighe PJ, Bihorac A, Rashidi P. Deep EHR: A Survey of Recent Advances in Deep Learning Techniques for Electronic Health Record (EHR) Analysis. IEEE J Biomed Health Inform. 2018 Sep;22(5):1589–604.

14. Dunn J, Runge R, Snyder M. Wearables and the medical revolution. Pers Med. 2018 Sep;15(5):429–48.

15. Meng Y, Speier W, Ong MK, Arnold CW. Bidirectional Representation Learning from Transformers using Multimodal Electronic Health Record Data to Predict Depression. IEEE J Biomed Health Inform. 2021 Aug;25(8):3121–9.

16. Shao Y, Cheng Y, Nelson SJ, Kokkinos P, Zamrini EY, Ahmed A, et al. Hybrid Value-Aware Transformer Architecture for Joint Learning from Longitudinal and Non-Longitudinal Clinical Data [Internet]. Health Informatics; 2023 Mar [cited 2023 May 27]. Available from: http://medrxiv.org/lookup/doi/10.1101/2023.03.09.23287046

17. van Timmeren JE, Cester D, Tanadini-Lang S, Alkadhi H, Baessler B. Radiomics in medical imaging—"how-to" guide and critical reflection. Insights Imaging. 2020 Dec;11(1):91.

18. Li J, Chen J, Tang Y, Wang C, Landman BA, Zhou SK. Transforming medical imaging with Transformers? A comparative review of key properties, current progresses, and future perspectives. Med Image Anal. 2023 Apr;85:102762.

19. The Sequence of the Human Genome | Science [Internet]. [cited 2025 May 17]. Available from: https://www.science.org/doi/10.1126/science.1058040?url_ver=Z39.88-2003&rfr_id=ori:rid:crossref.org&rfr_dat=cr_pub%20%200pubmed

20. Lewis CM, Vassos E. Polygenic risk scores: from research tools to clinical instruments. Genome Med. 2020 Dec;12(1):44.

21. Choi SW, Mak TSH, O'Reilly PF. Tutorial: a guide to performing polygenic risk score analyses. Nat Protoc. 2020 Sep 1;15(9):2759–72.

22. Dietz S, Altstidl T, Zanca D, Eskofier B, Nguyen A. How Intermodal Interaction Affects the Performance of Deep Multimodal Fusion for Mixed-Type Time Series [Internet]. arXiv; 2024 [cited 2025 May 17]. Available from: http://arxiv.org/abs/2406.15098

23. Zadeh A, Chen M, Poria S, Cambria E, Morency LP. Tensor Fusion Network for Multimodal Sentiment Analysis [Internet]. arXiv; 2017 [cited 2025 May 17]. Available from: http://arxiv.org/abs/1707.07250

24. An Y, Liu Y, Chen X, Sheng Y. TERTIAN: Clinical Endpoint Prediction in ICU via Time-Aware Transformer-Based Hierarchical Attention Network. Hošovský A, editor. Comput Intell Neurosci. 2022 Dec 16;2022:1–13.

25. The burden of chronic kidney disease in Australian patients with type 2 diabetes (the NEFRON study) - Thomas - 2006 - Medical Journal of Australia - Wiley Online Library [Internet]. [cited 2025 May 21]. Available from: https://onlinelibrary.wiley.com/doi/abs/10.5694/j.1326-5377.2006.tb00499.x

26. Thomas MC, Brownlee M, Susztak K, Sharma K, Jandeleit-Dahm KAM, Zoungas S, et al. Diabetic kidney disease. Nat Rev Dis Primer. 2015 Jul 30;1(1):15018.

27. Ueki K, Sasako T, Okazaki Y, Miyake K, Nangaku M, Ohashi Y, et al. Multifactorial intervention has a significant effect on diabetic kidney disease in patients with type 2 diabetes. Kidney Int. 2021 Jan 1;99(1):256–66.

28. Sandbæk A, Griffin SJ, Sharp SJ, Simmons RK, Borch-Johnsen K, Rutten GEHM, et al. Effect of Early Multifactorial Therapy Compared With Routine Care on Microvascular Outcomes at 5 Years in People With Screen-Detected Diabetes: A Randomized Controlled Trial: The ADDITION-Europe Study. Diabetes Care. 2014 Jun 12;37(7):2015–23.

29. Nelson RG, Grams ME, Ballew SH, Sang Y, Azizi F, Chadban SJ, et al. Development of Risk Prediction Equations for Incident Chronic Kidney Disease. JAMA. 2019 Dec 3;322(21):2104–14.

30. Li G, Li J, Tian F, Ren J, Guo Z, Pan S, et al. A 10-year retrospective cohort of diabetic patients in a large medical institution: Utilizing multiple machine learning models for diabetic kidney disease prediction. Digit Health. 2024 Jul 21;10:20552076241265220.

31. Li Y, Jin N, Zhan Q, Huang Y, Sun A, Yin F, et al. Machine learning-based risk predictive models for diabetic kidney disease in type 2 diabetes mellitus patients: a systematic review and meta-analysis. Front Endocrinol. 2025 Mar 3;16:1495306.

32. Chong YH, Fan Q, Tham YC, Gan A, Tan SP, Tan G, et al. Type 2 Diabetes Genetic Variants and Risk of Diabetic Retinopathy. Ophthalmology. 2017 Mar 1;124(3):336–42.

33. Ali AS, Pham C, Morahan G, Ekinci EI. Genetic Risk Scores Identify People at High Risk of Developing Diabetic Kidney Disease: A Systematic Review. J Clin Endocrinol Metab. 2023 Dec 1;109(5):1189–97.

34. Cole JB, Florez JC. Genetics of diabetes mellitus and diabetes complications. Nat Rev Nephrol. 2020 Jul;16(7):377–90.

35. van Zuydam NR, Ahlqvist E, Sandholm N, Deshmukh H, Rayner NW, Abdalla M, et al. A Genome-Wide Association Study of Diabetic Kidney Disease in Subjects With Type 2 Diabetes. Diabetes. 2018 Jul;67(7):1414–27.

36. Khan A, Turchin MC, Patki A, Srinivasasainagendra V, Shang N, Nadukuru R, et al. Genome-wide polygenic score to predict chronic kidney disease across ancestries. Nat Med. 2022 Jul 1;28(7):1412–20.

37. Madan S, Lentzen M, Brandt J, Rueckert D, Hofmann-Apitius M, Fröhlich H. Transformer models in biomedicine. BMC Med Inform Decis Mak. 2024 Jul 29;24(1):214.

38. Placido D, Yuan B, Hjaltelin JX, Zheng C, Haue AD, Chmura PJ, et al. A deep learning algorithm to predict risk of pancreatic cancer from disease trajectories. Nat Med. 2023 May;29(5):1113–22.

39. Li Y, Rao S, Solares JRA, Hassaine A, Ramakrishnan R, Canoy D, et al. BEHRT: Transformer for Electronic Health Records. Sci Rep. 2020 Apr 28;10(1):7155.

40. Li Y, Mamouei M, Salimi-Khorshidi G, Rao S, Hassaine A, Canoy D, et al. Hi-BEHRT: Hierarchical Transformer-Based Model for Accurate Prediction of Clinical Events Using Multimodal Longitudinal Electronic Health Records. IEEE J Biomed Health Inform. 2023 Feb;27(2):1106–17.

41. Zhou HY, Yu Y, Wang C, Zhang S, Gao Y, Pan J, et al. A transformer-based representation-learning model with unified processing of multimodal input for clinical diagnostics. Nat Biomed Eng. 2023 Jun;7(6):743–55.

42. LOINC, a Universal Standard for Identifying Laboratory Observations: A 5-Year Update | Clinical Chemistry | Oxford Academic [Internet]. [cited 2025 May 1]. Available from: https://academic.oup.com/clinchem/article-abstract/49/4/624/5641953?redirectedFrom=fulltext&login=false

43. Lambert SA, Gil L, Jupp S, Ritchie SC, Xu Y, Buniello A, et al. The Polygenic Score Catalog as an open database for reproducibility and systematic evaluation. Nat Genet. 2021 Apr;53(4):420–5.

44. Lambert SA, Wingfield B, Gibson JT, Gil L, Ramachandran S, Yvon F, et al. Enhancing the Polygenic Score Catalog with tools for score calculation and ancestry normalization. Nat Genet. 2024 Oct;56(10):1989–94.

45. Tanigawa Y, Qian J, Venkataraman G, Justesen JM, Li R, Tibshirani R, et al. Significant sparse polygenic risk scores across 813 traits in UK Biobank. PLoS Genet. 2022 Mar 24;18(3):e1010105.

46. Genovese G, Rockweiler NB, Gorman BR, Bigdeli TB, Pato MT, Pato CN, et al. BCFtools/liftover: an accurate and comprehensive tool to convert genetic variants across genome assemblies. Bioinformatics. 2024 Feb 1;40(2):btae038.

47. Landrum MJ, Lee JM, Riley GR, Jang W, Rubinstein WS, Church DM, et al. ClinVar: public archive of relationships among sequence variation and human phenotype. Nucleic Acids Res. 2014 Jan 1;42(Database issue):D980–5.

48. Tesfaye S, Cronin RM, Lopez-Class M, Chen Q, Foster CS, Gu CA, et al. Measuring social determinants of health in the All of Us Research Program. Sci Rep. 2024 Apr 16;14(1):8815.

49. Guide A, Garbett S, Feng X, Mapes BM, Cook J, Sulieman L, et al. Balancing efficacy and computational burden: weighted mean, multiple imputation, and inverse probability weighting methods for item non-response in reliable scales. J Am Med Inform Assoc. 2024 Dec 1;31(12):2869–79.

50. Bauman A, Bull F, Chey T, Craig CL, Ainsworth BE, Sallis JF, et al. The International Prevalence Study on Physical Activity: results from 20 countries. Int J Behav Nutr Phys Act. 2009 Mar 31;6:21.

51. Liu Y, Ott M, Goyal N, Du J, Joshi M, Chen D, et al. RoBERTa: A Robustly Optimized BERT Pretraining Approach [Internet]. arXiv; 2019 [cited 2025 May 1]. Available from: http://arxiv.org/abs/1907.11692

52. Sbaraglia M, Bellan E, Dei Tos AP. The 2020 WHO Classification of Soft Tissue Tumours: news and perspectives. Pathologica. 2020 Nov;113(2):70–84.

53. Clark MA, Fisher C, Judson I, Thomas JM. Soft-Tissue Sarcomas in Adults. N Engl J Med. 2005 Aug 18;353(7):701–11.

54. Casali PG, Abecassis N, Bauer S, Biagini R, Bielack S, Bonvalot S, et al. Soft tissue and visceral sarcomas: ESMO–EURACAN Clinical Practice Guidelines for diagnosis, treatment and follow-up. Ann Oncol. 2018 Oct;29:iv51–67.

55. Local relapse of soft tissue sarcoma of the extremities or trunk wall operated on with wide margins without radiation therapy - PMC [Internet]. [cited 2025 Mar 4]. Available from: https://pmc.ncbi.nlm.nih.gov/articles/PMC10144696/

56. Muehlhofer HML, Schlossmacher B, Lenze U, Lenze F, Burgkart R, Gersing AS, et al. Oncological Outcome and Prognostic Factors of Surgery for Soft Tissue Sarcoma After Neoadjuvant or Adjuvant Radiation Therapy: A Retrospective Analysis over 15 Years. Anticancer Res. 2021 Jan 1;41(1):359–68.

57. Bilgeri A, Klein A, Lindner LH, Nachbichler S, Knösel T, Birkenmaier C, et al. The Effect of Resection Margin on Local Recurrence and Survival in High Grade Soft Tissue Sarcoma of the Extremities: How Far Is Far Enough? Cancers. 2020 Sep 8;12(9):2560.

58. Jang WY, Kim HS, Han I. Impact of surgical margin on survival in extremity soft tissue sarcoma. Medicine (Baltimore). 2021 Jan 22;100(3):e24124.

59. Saxby NE, An Q, Miller BJ. Local Recurrence of Soft Tissue Sarcoma Revisited: Is there a Role for “Selective” Radiation? Iowa Orthop J. 2022 Jun;42(1):239–48.

60. Fujiwara T, Stevenson J, Parry M, Tsuda Y, Kaneuchi Y, Jeys L. The adequacy of resection margin for non-infiltrative soft-tissue sarcomas. Eur J Surg Oncol. 2021 Feb 1;47(2):429–35.

61. Yurtbay A, Aydın Şimşek Ş, Cengiz T, Bariş YS, Say F, Dabak N. The Impact of Surgical Margin Distance on Local Recurrence and Survival in Patients with Soft Tissue Sarcoma. Medicina (Mex). 2025 Feb;61(2):289.

62. Chen CC, Wu YY, Kao JT, Chang C, Huang SC, Shih H. Impact of resection margin on outcome in soft-tissue sarcomas of the extremities treated with limb-sparing surgery and postoperative radiotherapy. World J Surg Oncol. 2024 Apr 26;22:113.

63. Montreuil J, Kholodovsky E, Markowitz M, Torralbas Fitz S, Campano D, Erik Geiger J, et al. Rethinking tumor viability as prognostic factor in soft tissue sarcoma. J Orthop. 2025 Oct 1;68:7–14.

64. Sambri A, Caldari E, Fiore M, Zucchini R, Giannini C, Pirini MG, et al. Margin Assessment in Soft Tissue Sarcomas: Review of the Literature. Cancers. 2021 Jan;13(7):1687.

65. Peeken JC, Neumann J, Asadpour R, Leonhardt Y, Moreira JR, Hippe DS, et al. Prognostic Assessment in High-Grade Soft-Tissue Sarcoma Patients: A Comparison of Semantic Image Analysis and Radiomics. Cancers. 2021 Apr 16;13(8):1929.

66. Nakamura T, Matsumine A, Matsubara T, Asanuma K, Yada Y, Hagi T, et al. Infiltrative tumor growth patterns on magnetic resonance imaging associated with systemic inflammation and oncological outcome in patients with high-grade soft-tissue sarcoma. PLOS ONE. 2017 Jul 20;12(7):e0181787.

67. Tsukamoto S, Mavrogenis AF, Tanaka Y, Errani C. Imaging of Soft Tissue Tumors. Curr Med Imaging. 17(2):197–216.

68. Crombé A, Marcellin PJ, Buy X, Stoeckle E, Brouste V, Italiano A, et al. Soft-Tissue Sarcomas: Assessment of MRI Features Correlating with Histologic Grade and Patient Outcome. Radiology. 2019 Jun;291(3):710–21.

69. Lee S, Jung JY, Nam Y, Jung CK, Lee SY, Lee J, et al. Diagnosis of Marginal Infiltration in Soft Tissue Sarcoma by Radiomics Approach Using T2-Weighted Dixon Sequence. J Magn Reson Imaging. 2023;57(3):752–60.

70. Checklist for Artificial Intelligence in Medical Imaging (CLAIM): A Guide for Authors and Reviewers [Internet]. [cited 2023 Sep 14]. Available from: https://pubs.rsna.org/doi/epdf/10.1148/ryai.2020200029

71. DeLong ER, DeLong DM, Clarke-Pearson DL. Comparing the areas under two or more correlated receiver operating characteristic curves: a nonparametric approach. Biometrics. 1988 Sep;44(3):837–45.

72. Stacchiotti S, Verderio P, Messina A, Morosi C, Collini P, Llombart-Bosch A, et al. Tumor response assessment by modified Choi criteria in localized high-risk soft tissue sarcoma treated with chemotherapy. Cancer. 2012;118(23):5857–66.

73. Eisenhauer EA, Therasse P, Bogaerts J, Schwartz LH, Sargent D, Ford R, et al. New response evaluation criteria in solid tumours: Revised RECIST guideline (version 1.1). Eur J Cancer. 2009 Jan;45(2):228–47.

74. Clinical significance of the tail-like pattern in soft-tissue sarcomas on magnetic resonance imaging - ScienceDirect [Internet]. [cited 2025 Mar 11]. Available from: https://www.sciencedirect.com/science/article/abs/pii/S0949265818301623?via%3Dihub

75. Peeken JC, Asadpour R, Specht K, Chen EY, Klymenko O, Akinkuoroye V, et al. MRI-based delta-radiomics predicts pathologic complete response in high-grade soft-tissue sarcoma patients treated with neoadjuvant therapy. Radiother Oncol J Eur Soc Ther Radiol Oncol. 2021 Nov;164:73–82.

76. Peeken JC, Spraker MB, Knebel C, Dapper H, Pfeiffer D, Devecka M, et al. Tumor grading of soft tissue sarcomas using MRI-based radiomics. EBioMedicine. 2019 Oct;48:332–40.

77. Peeken JC, Asadpour R, Specht K, Chen EY, Klymenko O, Akinkuoroye V, et al. MRI-based delta-radiomics predicts pathologic complete response in high-grade soft-tissue sarcoma patients treated with neoadjuvant therapy. Radiother Oncol J Eur Soc Ther Radiol Oncol. 2021 Nov;164:73–82.

78. Navarro F, Dapper H, Asadpour R, Knebel C, Spraker MB, Schwarze V, et al. Development and External Validation of Deep-Learning-Based Tumor Grading Models in Soft-Tissue Sarcoma Patients Using MR Imaging. Cancers. 2021 Jun 8;13(12):2866.

79. Spraker MB, Wootton LS, Hippe DS, Ball KC, Peeken JC, Macomber MW, et al. MRI Radiomic Features Are Independently Associated With Overall Survival in Soft Tissue Sarcoma. Adv Radiat Oncol. 2019 Apr;4(2):413–21.

80. Cahan N, Klang E, Marom EM, Soffer S, Barash Y, Burshtein E, et al. Multimodal fusion models for pulmonary embolism mortality prediction. Sci Rep. 2023 May 9;13(1):7544.

81. Iacobini C, Pugliese G, Blasetti Fantauzzi C, Federici M, Menini S. Metabolically healthy versus metabolically unhealthy obesity. Metabolism. 2019 Mar 1;92:51–60.

82. Després JP, Lemieux I. Abdominal obesity and metabolic syndrome. Nature. 2006 Dec;444(7121):881–7.

83. Bradshaw PT. Body composition and cancer survival: a narrative review. Br J Cancer. 2023 Oct 27;130(2):176.

84. Santhanam P, Nath T, Peng C, Bai H, Zhang H, Ahima RS, et al. Artificial intelligence and body composition. Diabetes Metab Syndr. 2023 Mar;17(3):102732.

85. Full article: Pathogenic potential of adipose tissue and metabolic consequences of adipocyte hypertrophy and increased visceral adiposity [Internet]. [cited 2024 Dec 6]. Available from: https://www.tandfonline.com/doi/full/10.1586/14779072.6.3.343

86. Pickhardt PJ, Graffy PM, Perez AA, Lubner MG, Elton DC, Summers RM. Opportunistic Screening at Abdominal CT: Use of Automated Body Composition Biomarkers for Added Cardiometabolic Value. Radiogr Rev Publ Radiol Soc N Am Inc. 2021;41(2):524–42.

87. Linder N, Denecke T, Busse H. Body composition analysis by radiological imaging - methods, applications, and prospects. ROFO Fortschr Geb Rontgenstr Nuklearmed. 2024 Oct;196(10):1046–54.

88. Magudia K, Bridge CP, Bay CP, Babic A, Fintelmann FJ, Troschel FM, et al. Population-Scale CT-based Body Composition Analysis of a Large Outpatient Population Using Deep Learning to Derive Age-, Sex-, and Race-specific Reference Curves. Radiology. 2021 Feb;298(2):319–29.

89. Mourtzakis M, Prado CMM, Lieffers JR, Reiman T, McCargar LJ, Baracos VE. A practical and precise approach to quantification of body composition in cancer patients using computed tomography images acquired during routine care. Appl Physiol Nutr Metab. 2008 Oct;33(5):997–1006.

90. Pyrros A, Borstelmann SM, Mantravadi R, Zaiman Z, Thomas K, Price B, et al. Opportunistic detection of type 2 diabetes using deep learning from frontal chest radiographs. Nat Commun. 2023 Jul 7;14(1):4039.

91. Wasserthal J, Breit HC, Meyer MT, Pradella M, Hinck D, Sauter AW, et al. TotalSegmentator: Robust Segmentation of 104 Anatomic Structures in CT Images. Radiol Artif Intell [Internet]. 2023 Sep [cited 2024 Jun 19];5(5). Available from: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC10546353/

92. Blankemeier L, Desai A, Chaves JMZ, Wentland A, Yao S, Reis E, et al. Comp2Comp: Open-Source Body Composition Assessment on Computed Tomography [Internet]. arXiv; 2023 [cited 2024 Oct 9]. Available from: http://arxiv.org/abs/2302.06568

93. Agatston AS, Janowitz WR, Hildner FJ, Zusmer NR, Viamonte M, Detrano R. Quantification of coronary artery calcium using ultrafast computed tomography. J Am Coll Cardiol. 1990 Mar 15;15(4):827–32.

94. A Deep Learning Algorithm for Automatic 3D Segmentation of Rotator Cuff Muscle and Fat from Clinical MRI Scans [Internet]. [cited 2023 May 16]. Available from: https://pubs.rsna.org/doi/epdf/10.1148/ryai.220132

95. He K, Zhang X, Ren S, Sun J. Deep Residual Learning for Image Recognition [Internet]. arXiv; 2015 [cited 2024 Oct 9]. Available from: http://arxiv.org/abs/1512.03385

96. Huang SC, Pareek A, Zamanian R, Banerjee I, Lungren MP. Multimodal fusion with deep neural networks for leveraging CT imaging and electronic health record: a case-study in pulmonary embolism detection. Sci Rep. 2020 Dec 17;10(1):22147.

97. Gokcesu K, Gokcesu H. Generalized Huber Loss for Robust Learning and its Efficient Minimization for a Robust Statistics [Internet]. arXiv; 2021 [cited 2024 Oct 9]. Available from: http://arxiv.org/abs/2108.12627

98. van der Velden BHM, Kuijf HJ, Gilhuijs KGA, Viergever MA. Explainable artificial intelligence (XAI) in deep learning-based medical image analysis. Med Image Anal. 2022 Jul 1;79:102470.

99. Tagliafico AS, Bignotti B, Torri L, Rossi F. Sarcopenia: how to measure, when and why. Radiol Med (Torino). 2022 Jan 18;127(3):228.