

Causal inference in radiomics: framework, mechanisms, and algorithms

Debashis Ghosh*

*Departments of Biostatistics and Informatics, Colorado School of Public Health

debashis.ghosh@cuanschutz.edu

Emily Mastej†

† Computational Biosciences Program, University of Colorado Anschutz Medical Campus

emily.mastej@cuanschutz.edu

Rajan Jain‡

Departments of Radiology and Neurosurgery

NYU Langone Medical Center

rajan.jain@nyulangone.org

and

Yoon-Seong Choi**

Department of Radiology

Yonsei University College of Medicine

yoonseong.choi07@gmail.com

Summary

The widespread use of machine learning algorithms in radiomics has led to a proliferation of flexible prognostic models for clinical outcomes. However, a limitation of these techniques is their black-box nature, which prevents the ability for increased mechanistic phenomenological understanding. In this article, we develop an inferential framework for estimating causal effects with radiomics data. A new challenge is that the exposure of interest is latent so that new estimation procedures are needed. We leverage a multivariate version of partial least squares for causal effect estimation. The methodology is illustrated with applications to two radiomics datasets, one in osteosarcoma and one in glioblastoma.

Keywords: Latent causal effect; link-free inference; Medical imaging; personalized medicine.

1 Introduction

Radiomics explores relationships between image-derived characteristics of a tumor and other parameters, including clinical outcomes and genomic profiles, including gene expression, somatic mutations, and DNA methylation [1]. In particular, several groups have built classifiers to predict tumor molecular phenotypes using radiomic inputs (e.g., [2, 3, 4, 5]). More recently, there has been tremendous interest in using modern machine learning and in particular deep learning tools in order to build state-of-the-art classifiers for predictions [6, 7, 8].

In spite of their state-of-the-art performance, use of these complex models comes at a cost. Because many of these classifiers are “black-box” in nature, clinicians consequently have a difficult time understanding the predictions. More generally, most work in radiomics has focused on pattern-based associations in the data with a machine learning viewpoint in one of two ways. First, these analyses could take the form of clustering algorithms, such as t-SNE or UMAP, in which interesting clusterings lead to followup discoveries. Second, a classification or supervised learning framework could be adopted in which the radiomics features could be used to predict a class label or phenotype of interest. For this approach, a typical evaluation is classification accuracy of the ROC curve or F1 values where higher values are better. While there has been tremendous successes in radiomics with these machine learning techniques, it still remains elusive from the end goal of developing mechanistic insights into tumorigenesis.

In this article, we seek to introduce causal modelling concepts into radiomics. While there has been much work on using these ideas in genomics (e.g., [9, 10]) and brain imaging (e.g., [11]), their application to radiomics has not occurred. We argue that adopting this viewpoint in radiomics has the following advantages:

1. It allows one to view radiomics as measurements of properties of the tumor and its characteristics.
2. Developing a causal model allows one to link tumorigenic mechanisms to observed data.
3. The causal inferential pathway is compatible with the trend towards systems biology [12] while not being as purely reductionist as approaches such as those based on mathematical models or ordinary differential equations.

However, a challenge that this approach introduces is that we must view the tumor as a latent construct, and causal inference with latent structures is much more challenging. There has been much recent interest in the use of latent class modelling of treatment effects on outcomes [13]. Bandeen-Roche et al. [14] proposed a latent class modelling approach in

which pseudodraws from the inferred latent class distribution are then used to fit regression models on the outcome. Multiple pseudodraws are generated, and the multiple regression results are combined using Rubin’s imputation rules [15]. A simpler approach is to use a classify/analyze approach [16] in which each individual is assigned to a latent class, and then the group assignment is used as a covariate for which standard propensity score methods can be applied. In a recent study, Schuler et al. [17] compared these approaches, along with a joint modelling approach developed by Kang and Schafer from an unpublished technical report at the Methodology Center of Penn State University. The approaches can be broadly grouped as being 1-step versus 3-step approaches [18]. The former methods fit a joint model describing both the latent classes as well as the latent classes’ effect on the outcome. The latter go through a series of three steps: (a) fit a model to describe the joint classes; (b) assign membership of the individuals to these classes; (c) fit a regression model of the outcome on the inferred latent class. Schuler et al. [17] provide a nice discussion of the strengths and weaknesses of each approach. In their conclusions, they suggest that 1-step methods offer many advantages but that one barrier to their implementation is computational. To be precise, estimation based on a joint likelihood for both the latent class and causal effect modelling might have issues with numerical convergence.

Our new contributions to the literature in the current paper are the following:

1. Extension and formalization of the classical potential outcomes framework [19, 20] to accomodate latent treatment effects. This entails developing the necessary assumptions for definition and identification of an appropriate causal estimand with observed data. This leads to a new quantity, the local latent average causal effect.
2. Development of a new estimation procedure for the local causal effect with latent variables. This leverages techniques from sufficient dimension reduction and partial least squares [21].

To our knowledge, the use of latent causal inferential techniques has not been developed for effect estimation in radiomics or genomics. The most related technique comes from genetics, where principal components-type approaches are used to adjust for population stratification, which is a type of confounding [22, 23]. However, in that setup, latent variables are used to model confounders, not the main effect of interest, which is the focus of the current article. As a proof of concept, we apply our approach to two radiomics datasets in the literature, one from glioblastoma, the other a public available data from osteosarcoma [24].

2 Motivating Datasets:

In this article, we will use two datasets to illustrate the methodology as a proof of concept. We use these because the distribution of the outcome is different. For the study described in Section 2.1., the outcome is continuous, while for that in Section 2.2., the outcome is binary.

2.1 Glioblastoma multiforme study

Glioma is the most common type of brain cancer; it develops in the glial cells [25]. Among these, glioblastoma multiforme (GBM) is the most frequent and malignant histologic type. Patients with GBM have on average 3% 5-year survival after diagnosis [25]. The dataset we work with comes from the Cancer Genome Atlas, consisting of data on 226 subjects with GBM. For these subjects, imaging was done using three protocols: T1, T2 and FLAIR. In this paper, we only focus on the first two. T1 and T2 refer to protocols that utilize two different properties of fMRI. With fMRI, the magnetic current induces a magnetic field, and T1 refers to the speed at which the electron spins in the blood realign with the recovery of the longitudinal orientation. T2 refers to the loss of magnetization as a result of the loss in phase coherence of the electrons. The T1 and T2 images were represented in DICOM format, which was then converted to NIFTI format and processed for standardization with 1mm isovoxel resolution as follows:

1. Postcontrast T1-weighted images (T1C) were resampled to 1mm isovoxel resolution;
2. T2 images were registered to T1C images after skull stripping, using the FMRIB software library (<http://fsl.fmrib.ox.ac.uk/fsl/fslwiki/FSL>).
3. Image signal intensity was normalized using the WhiteStripe R package.

The tumor areas, defined as areas of T2 hyper-intense tumor and edema on FLAIR images, were segmented by using semi-automatic methods, including signal intensity thresholding, region growing, and edge detection, with an open source software (Medical Image Processing, Analysis and Visualization, <https://mipav.cit.nih.gov/>). Radiomic features were extracted from all isotropic voxel image segmented regions of interest (ROIs) using pyRadiomics [26]. Extraction settings were configured to features from original images, as well as Wavelet filtered and Laplacian of Gaussian (LoG) filtered images and were calculated considering adjacent voxel in 3 dimensions. In total, 1046 radiomic features were extracted per image.

2.2 Osteosarcoma study

Osteosarcoma is a cancer that usually develops in the cells that form bone, the osteoclasts. It happens most often in children, adolescents, and young adults. In a recent study, Zhang et al. [27] conducted a study of 102 subjects with osteosarcoma who underwent neoadjuvant chemotherapy. Prior to having treatment, they received a dynamic contrast-enhanced MRI (DCE-MRI) scan. The Response Evaluation Criteria in Solid Tumors were used to evaluate the neoadjuvant chemotherapy response as effective (complete remission and/or partial remission) or ineffective (stable and progressive disease).

Using the Radcloud software platform, Zhang et al. [27] extracted a total of 1,409 quantitative imaging features. They can be divided into four groups: (a) Group 1 represent typical summaries for the distribution of voxel intensities within the MR image; (b) Group 2 are three-dimensional features that reflect the shape and size of the region; (c) Group 3 are second-order texture features that quantify region heterogeneity differences, calculated from gray-level run length and gray-level co-occurrence texture matrices; (d) Group 4 contains 1,302 first-order statistics and texture features after applying Laplacian, logarithmic, exponential, and wavelet filters on the image. The goal is to see whether or not radiomics can predict treatment response.

3 Proposed Methodology

3.1 Potential outcomes framework

We first review the potential outcomes framework of Rubin [19] and Holland [20] and begin by assuming that the treatment is observed. For the sake of exposition, we will assume that it is continuous, similar to Imai and van Dyk [28]. Let Z denote the treatment, with possible values $z \in \mathcal{Z}$. Define $\{Y_i(z) : z \in \mathcal{Z}\}$ to the set of potential outcomes for subject i , $i = 1, \dots, n$; $Y_i(z)$ represents the potential outcome for subject i with the treatment equals z . In addition, we assume the existence of a set of confounders $\mathbf{X}$. For proper causal inference within the potential outcomes framework, one needs the assumption based on strong ignorability of the treatment:

$$\{Y(z) : z \in \mathcal{Z}\} \perp\!\!\!\perp Z | \mathbf{X}, \quad (1)$$

which in words states that treatment assignment is conditionally independent of the set of potential outcomes given covariates. In the case of binary treatment, Rosenbaum and Rubin [29] refer to (1) as the strongly ignorable treatment assumption. Heuristically, what (1) implies is that the potential outcomes can be viewed as predefined random variables. The randomness in the populations occurs due to the nonignorable missing data mechanism, in

the sense of Little and Rubin [15], that makes only one of the potential outcomes observable for each subject. In addition, we make the assumption of consistency so that the the observed outcome for a subject coincides with the corresponding potential outcome.

Based on the potential outcomes, we can define the following local causal effect parameter

$$LCE_i(z) = Y_i(z + 1) - Y_i(z), \quad z \in \mathcal{Z} \quad (2)$$

for $i = 1, \dots, n$. We also note the dependence of (2) on the treatment. If we assume that the effect is constant over levels of Z , then it would be possible to pool effects to result in a statistically more efficient estimator of the local causal effect.

Two more assumptions that are commonly invoked in the causal inference literature are the positivity assumption and the common support condition. The former states that $E(Z|\mathbf{X}) \neq 0$ for all possible values of $\mathbf{X}$. In other words, there exist no regions of the confounder distribution that preclude observe any possible value of the treatment. The common support condition states that there is sufficient overlap in $\mathbf{X}$ across all values of Z .

If we were to average the local causal effects (2) over all subjects ($i = 1, \dots, n$), then this would correspond to a *local* average causal effect. In the case where the treatment is binary, this effect reduces to the average causal effect that has been considered in the literature.

Analogous to the propensity score of Rosenbaum and Rubin [29] in the case of binary treatment, we can consider a quantity representing the conditional mean of treatment given confounders, $E(Z|\mathbf{X})$. This is a special case of the generalized propensity score of Imai and van Dyk [28] and has several desirable properties. The first is that it reduces the modelling of confounders to modelling a conditional mean of treatment, which reduces the dimension. Second, provided $E(Z|\mathbf{X})$ is correctly modelled, then it functions as a balancing score in that if (1) holds, then $\{Y(z) : z \in \mathcal{Z}\} \perp\!\!\!\perp Z|E(Z|\mathbf{X})$. The properties of $E(Z|\mathbf{X})$ lead to a natural strategy for performing causal inference:

1. Fit a model for $E(Z|\mathbf{X})$;
2. Given the predicted conditional mean in (1), perform a regression of Y on Z in which an adjustment is made using $E(Z|\mathbf{X})$.

There is a variety of approaches to perform adjustment in the second step of this algorithm. These include inverse weighting methods, regression adjustment, matching, subclassification or a combination thereof. Please see Lunceford and Davidian [30] for more discussion on these methods.

3.2 Latent treatment: model formulation and estimation

We now relax the assumption that the treatment Z be observed. Instead, we now have available several observations $\mathbf{U} \equiv (U_1, \dots, U_K)$ that capture the latent treatment variable Z . A commonly used assumption here is the so-called local independence assumption [31], which states that conditional on Z , $U_1, \dots, U_K$ are conditionally independent. We can then factorize the joint distribution of the potential outcomes (denoted here as $Y(\cdot)$) Z , $\mathbf{X}$ and $\mathbf{U}$ as

$$\begin{aligned} f_{Y(\cdot), Z, \mathbf{X}, \mathbf{U}} &= f_{Y(\cdot)|Z, \mathbf{X}, \mathbf{U}} f_{Z|\mathbf{X}, \mathbf{U}} f_{\mathbf{X}, \mathbf{U}} \\ &= f_{Y(\cdot)|\mathbf{U}, \mathbf{X}} f_{Z|\mathbf{X}, \mathbf{U}} f_{\mathbf{X}, \mathbf{U}}, \end{aligned}$$

where we have the assumption (1) to simplify the conditional distribution going from the first line to the second. If we further assumption that the joint density $f_{\mathbf{X}, \mathbf{U}}$ are ancillary for the causal parameters of interest, then we can base our approach to inference by specifying likelihoods corresponding to $f_{Y(\cdot)|\mathbf{U}, \mathbf{X}}$ and $f_{Z|\mathbf{X}, \mathbf{U}}$, respectively. Recapping, here are the assumptions we need for valid causal effect estimation of the LCE in the latent case:

1. $Z \perp\!\!\!\perp \{Y(z) : z \in \mathcal{Z}\}$;
2. The components of $\mathbf{U}$ are conditionally independent given Z ;
3. The components of $\mathbf{U}$ are conditionally independent of any variable S given Z ;
4. $E(Z|\mathbf{X})$ exists for all $\mathbf{X}$.

We note that Assumptions 2 and 3 look similar but are conceptually different. The former deals with the radiomics features providing conditionally independent information given the tumor and is referred to as local independence. By contrast, the latter has to do with measurement invariance [32], namely that the radiomics measurements are capturing the same concept independent of other variables. In fact, it is assumption 3 that is a very important one if one wishes to have any chance of the radiomics data analyses reflecting potentially generalizable findings.

Figure 1 depicts our conceptual model.

Figure 1 about here.

We could then convert Figure 1 into a causal diagram [33] which leverages the graphical model for causation [34]. We can use the traditional rules about directed acyclic graphs to model conditional independence. In particular, the observed radiomics data will be conditionally

independent given the latent variable. In addition, the latent variable d-separates [34] the confounders from $\mathbf{U}$. The causal effect we focus on in Figure 1 is that from the latent variable to the clinical outcome. This separation of the scientific estimands (i.e., causal effects) from the data represent one of the appealing features of Figure 1.

We note that the structure of Figure 1 is related to diagrams used by practitioners of structural equations modelling (SEM, [35]). In that literature, relationships between latent variables are referred to as the structural model, while those relating latent to measured are called measurement models. SEM combines the two types of models in order to induce a joint distribution for the observed data which is then used for estimation and inference. As described by Bollen and Pearl [35], the structural model is consistent with the potential outcomes framework that we outlined in §3.1.

Thinking of the radiomics data as the main effect in a causal analysis is consistent with a tumor progression model in which the cancer’s behavior at clinical diagnosis matters, and any biological preceding events can only play the role of confounders. Biologically, this means that confounders increase the propensity for tumorigenesis to occur.

4 Methodology

4.1 Proposition and partial least squares

Based on the assumptions and conditional independence statements we have laid out in the previous sections, we have following proposition.

Proposition 1. The random variable Z d-separates both $\mathbf{U}$ and $\mathbf{X}$ and Y and $\mathbf{X}$.

While proposition 1 is quite simple in nature, it in fact reveals a powerful result and leads to a new approach to causal effect estimation. In particular, we can estimate Z as a latent variable in two simultaneous regression models: (1) $\mathbf{U}$ on $\mathbf{X}$; (2) Y on $\mathbf{X}$. This simultaneous estimation has potential benefits largely due to the absence of direct arrows from $\mathbf{U}$ to Y in Figure 1. We have thus converted a causal effect estimation problem into a multi-task learning problem with a latent variable.

Our algorithms for causal effect estimation will be based on partial least squares (PLS) [36, 37]. PLS presumes that $\mathbf{U}_i$ and $\mathbf{X}_i$ ($i = 1, \dots, n$) are both linear functions of a set of common latent factors. An alternative characterization for PLS was given by Stone and Brooks [38]. Suppose that we are fitting the following model:

$$E(Y_i|\mathbf{Z}_i) = \mathbf{Z}_i^T \beta_0 \quad (3)$$

Then Stone and Brooks [38] consider the following class of objective functions:

$$\text{Var}(\mathbf{Z}_i^T \beta)^2 \text{Cov}(Y_i, \mathbf{Z}_i^T \beta)^{\alpha/(1-\alpha)-1},$$

where Var and Cov and short-hand notation for variance and covariance, and α is a number between 0 and 1. In this framework, values of $\alpha = 0$, $\alpha = 1/2$ and $\alpha = 1$ correspond to the objective functions maximized by ordinary least squares (OLS), partial least squares (PLS) and principal components regression (PCR), respectively. Using this framework, we find that the PLS algorithm acts as some hybrid between the usual least squares estimator and the principal components regression approach of Massy [39].

The algorithm for multivariate PLS that we will use is the kernel algorithm proposed by Dayal and MacGregor [40]. We assume that the columns of $\mathbf{U}$ and $\mathbf{X}$ are centered and scaled. Recall again that the PLS model formulation is given by

$$\begin{pmatrix} \mathbf{X} = \mathbf{T}\mathbf{P}' + \mathbf{E} \\ \mathbf{U} = \mathbf{V}\mathbf{Q}' + \mathbf{F} \end{pmatrix}, \quad (4)$$

where $\mathbf{T}$ and $\mathbf{V}$ are $n \times l$ matrices, and $\mathbf{P}$ and $\mathbf{Q}$ are the so-called loading matrices corresponding to $\mathbf{X}$ and $\mathbf{U}$, respectively. The dimension of $\mathbf{P}$ is $p \times l$, while for $\mathbf{Q}$, it is $m \times l$. The matrices $\mathbf{E}$ and $\mathbf{F}$ are the error terms with entries being independent and identically distributed normal random variables with mean zero and variance σ^2 .

The kernel algorithm proceeds as follows:

1. Compute the matrices $\mathbf{X}'\mathbf{X}$ and $\mathbf{X}'\mathbf{U}$.
2. Let $b = 1$. Compute $\mathbf{q}_1$ as the eigenvector corresponding to the largest eigenvalue of $\mathbf{U}'\mathbf{X}'\mathbf{X}\mathbf{U}$. Then set $\mathbf{w}'_b = (\mathbf{X}'\mathbf{U})_b\mathbf{q}_b$ and rescale the entries of $\mathbf{w}_b$ to have unit norm. For $b = 1$, $(\mathbf{X}'\mathbf{U})_b = \mathbf{X}'\mathbf{U}$; for $b > 1$, its definition will be given in (5).
3. Compute $\mathbf{r}_b$. For $b = 1$, $\mathbf{r}_1 = \mathbf{w}_1$, while for $b > 1$,

$$\mathbf{r}_b = \mathbf{w}_b - \sum_{a=1}^{b-1} \mathbf{p}'_a \mathbf{w}_b \mathbf{r}_a.$$

4. Compute $\mathbf{t}_b = \mathbf{X}\mathbf{r}_b$, $\mathbf{p}_b = \mathbf{t}'_b\mathbf{X}/\mathbf{t}'_b\mathbf{t}_b$ and $\mathbf{q}'_b = \mathbf{r}'_b(\mathbf{X}'\mathbf{U})/\mathbf{t}'_b\mathbf{t}_b$.
5. Compute

$$(\mathbf{X}'\mathbf{U})_{b+1} = (\mathbf{X}'\mathbf{U})_b - \mathbf{p}_b\mathbf{q}'_b(\mathbf{t}'_b\mathbf{t}_b) \quad (5)$$

6. Repeat steps (2)-(5).

At the end, the regression coefficients are given by the outer product of the matrix consisting of $\mathbf{r}_b$ and that consisting of $\mathbf{q}_b$. This kernel algorithm has been implemented in the `kernelpls.fit` function that is available in the `pls` package [41].

4.2 Theoretical justification

Partial least squares methods were given a justification for the single-index model by Naik and Tsai [21]. This was done by incorporating ideas from the field of sufficient dimension reduction [42]. This is a branch of statistics in which the goal is to develop ‘model-free’ procedures in order to summarize data while preserving regression relationships. The field started with the observation by Brillinger [43] in which ordinary least squares methods provide estimates that were consistent up to a sign for regression parameters in more general single-index models. A recent overview of the field can be found in Li [42].

We now develop a multivariate extension of the results of Naik and Tsai [21]. To do this, we consider a multivariate response for each subject that can be summarized as a K -dimensional vector $\mathbf{U}$ along with a p -dimensional vector of covariates $\mathbf{X}$.

We then formulate the following multivariate regression model:

$$\begin{pmatrix} U_1 \\ \vdots \\ U_K \end{pmatrix} = \begin{pmatrix} g_1(\beta'_1 \mathbf{X}, \epsilon_1) \\ \vdots \\ g_K(\beta'_K \mathbf{X}, \epsilon_K) \end{pmatrix} \quad (6)$$

where g_j ($j = 1, \dots, K$) are monotonic functions in both arguments and $\epsilon_1, \dots, \epsilon_K$ are random vectors representing the error distributions for the models. In (6), the p -dimensional parameter vectors $\beta_1, \dots, \beta_K$ specify the directions of interest. In the case where $K = 1$, model (6) reduces to a generalized single-index model. We make the following assumptions.

Assumption 1. $\mathbf{X}$ has a multivariate normal distribution.

Assumption 2. The covariance matrices $n^{-1}\mathbf{X}'\mathbf{X}$ and $n^{-1}\mathbf{X}'\mathbf{U}$ converge in probability to limits Σ_{xx} and Σ_{xu} , respectively.

Assumption 3. Taken as an operator, the range of Σ_{xx} coincides with the range of Σ_{xu} .

Based on these three assumptions, we have the following result.

Theorem: Under Assumptions 1–3, the multivariate PLS estimator converges in probability to a constant times $(\beta_1, \dots, \beta_m)$.

Proof: The multivariate PLS estimator can be expressed as $\hat{R}(\hat{R}'n^{-1}\mathbf{X}'\mathbf{X}\hat{R})^{-1}\hat{R}'n^{-1}\mathbf{X}'\mathbf{U}$, where $\hat{R}$ is the matrix derived from the Krylov sequence of matrices of $n^{-1}\mathbf{X}'\mathbf{X}$ and $n^{-1}\mathbf{X}'\mathbf{U}$. By assumption 2, this will converge to $R(R'\Sigma_{xx}R)^{-1}R'\Sigma_{xu}\beta^*$. By assumption 3, β^* will be in the space spanned by R so that $\Sigma_{xx}^{1/2}\beta^*$ will be in the space spanned by $\Sigma_{xx}^{1/2}R$. The statement of the theorem then follows.

One of the major assumptions in traditional sufficient dimension reduction procedures has been the linearity assumption, which is satisfied by multivariate normal distributions and more generally, elliptically symmetric distributions. This gets violated in situations with discrete predictors. Recently, Ghosh [44] has developed an interpretation of sufficient dimen-

sion reduction methods from an information-theoretic point of view. In this interpretation, the partial least squares algorithm can be viewed as an information-maximization operation under less restrictive distributional assumptions than those required in Theorem 1 of the paper.

4.3 Integration with causal modelling and implementation details

In examining Figure 1, we propose the following strategy for causal effect estimation.

1. Run a multivariate PLS regression of the radiomics data on the confounders.
2. Using the scores as the inferred latent treatment from the output of the PLS regression, perform causal inference of the treatment on outcome.

We have lots of choices on how to perform step 2. above. Our approach is to use regression adjustment as a means of inferring causal effects. For the standard error, we will use the nonparametric bootstrap [45].

We note that the outcome of the multivariate PLS can be fairly general. These include variables that are continuous, binary or unordered categorical. For right-censored failure time variables, we use the suggestion of Keles and Segal [46] and compute a first-stage martingale residual from a null model (i.e., one with no covariates). We then treat the residual as a continuous variable to be input into the partial least squares algorithm.

5 Numerical examples

5.1 Glioblastoma multiforme study

The confounders available in this analysis are gender, grade and IDH mutation status. Mutations of the isocitrate dehydrogenase (IDH1) gene has been shown to be a marker of oncogenesis and is one of the most specific biomarkers in the diagnostic classification of secondary GBM [47] Gliomas with mutated IDH have improved prognosis compared to gliomas with wild-type IDH and are detected by immunohistochemistry and magnetic resonance (MR) spectroscopy [48]. We first characterize survival in the population and by grade, IDH mutation status and gender. These are presented in Figure 2.

Figure 2 about here.

Based on the plots, we see that there are differences in survival based on grade and IDH mutation status and not for gender. We also note that 22 subjects are missing IDH mutation, while 1 is missing grade and gender.

We next applied the latent causal effect approach in the paper with two sets of confounders: (gender,grade) and (gender,grade,IDH mutation statu). We note that the former analysis will have 225 subjects, while the latter will have 203. The results are shown in Table 1.

Table 1 about here.

The model that is being fit is for the adjusted survival time as a function of the latent variable. Based on the analysis, we see that there is a highly significant effect of the treatment on outcome. Both analyses that higher values of the latent construct are associated with lower adjusted times to death.

5.2 Osteosarcoma study

In this dataset, there is only one confounder, the stage of cancer (stage IIB versus not). Of the 102 subjects, 75 were stage IIB. The effect of radiomics on treatment response was analyzed here. The multivariate PLS algorithm yields an average causal effect of 0.099, with an associated 95% confidence interval of $(-0.13, 0.20)$. As an alternative, we computed the first principal components using the radiomics data and fit a regression model of treatment response Y on the principal component and stage. Based on the fitted model, we obtained the following equation:

$$Y = 0.29 - 0.02PC1 + 0.23Stage,$$

with associated standard errors of 0.01 and 0.11 for the principal component and stage. While both variables are statistically significant at a 0.05 level of significance (p-values of 0.05 and 0.03 for PC score and stage, respectively), we note that the direction of the effect is reverse that from the multivariate PLS approach.

6 Discussion

In this article, we have introduced an approach to causal modelling with radiomics data. The assumptions needed for identification of the causal effects from observed data along with an estimation procedure for causal effects. We have demonstrated the application of the methodology to two radiomics datasets in cancer. Further evaluations are need to demonstrate validation of the methodology. To help readers, we have made code available to do the analysis with the osteosarcoma data at the following location:

<http://github.com/GhoshLab/CausalRadiomics>

An alternative approach to radiomics data is to treat it as a mediation variable in a causal effects analysis. There, it would be an intermediate variable, and genetic data would serve as the main exposure. Viewing the radiomics as mediation variable makes explicit the role of initiating events. In Figure 3, we suggest a DNA mutation as the beginning event, but other choices could be entertained.

Figure 3 about here.

In this setup, the radiomics is viewed as a downstream event, and the mutation will have effects on survival as mediated through the radiomics and effects outside the pathway. In [9], the authors used microRNA as the exposure and gene expression from several pathways as the mediators. For this setup, they develop tests of mediation and associated testing procedures. More recently, in [10], a Bayesian approach to mediation analysis as developed, in which methylation profiles were the high-dimensional mediators of a univariate exposure. Their algorithm was based on a variable selection procedure with shrinkage priors to select for mediators. Finally, we note the principal mediation directions approach of [11]. For their application, subject-level functional magnetic resonance imaging profiles were the mediator, and the goal was to understand the areas of the brain that mediated pain-invoked stimuli. In [11], the authors used a supervised principal components approach similar to what is presented here. We will explore extensions of our latent causal inference procedures to the mediation problem in future work.

Acknowledgments

The first author would like to acknowledge support from NIH R01 CA129102 and NSF DMS 1914937.

References

- [1] Maciej A Mazurowski. Radiogenomics: what it is and why it is important. *Journal of the American College of Radiology*, 12(8):862–866, 2015.
- [2] S. S. Yip, J. Kim, T. P. Coroller, C. Parmar, E. R. Velazquez, E. Huynh, R. H. Mak, and H. J. Aerts. Associations Between Somatic Mutations and Metabolic Imaging Phenotypes in Non-Small Cell Lung Cancer. *Journal of Nuclear Medicine*, 58(4):569–576, 04 2017.

- [3] P. Kickingereder, M. Gutz, J. Muschelli, A. Wick, U. Neuberger, R. T. Shinohara, M. Sill, M. Nowosielski, H. P. Schlemmer, A. Radbruch, W. Wick, M. Bendszus, K. H. Maier-Hein, and D. Bonekamp. Large-scale Radiomic Profiling of Recurrent Glioblastoma Identifies an Imaging Predictor for Stratifying Anti-Angiogenic Treatment Response. *Clin Cancer Res*, 22(23):5765–5771, Dec 2016.
- [4] E. Rios Velazquez, C. Parmar, Y. Liu, T. P. Coroller, G. Cruz, O. Stringfield, Z. Ye, M. Makrigiorgos, F. Fennessy, R. H. Mak, R. Gillies, J. Quackenbush, and H. J. W. L. Aerts. Somatic Mutations Drive Distinct Imaging Phenotypes in Lung Cancer. *Cancer Research*, 77(14):3922–3930, 07 2017.
- [5] Y. B. Xi, F. Guo, Z. L. Xu, C. Li, W. Wei, P. Tian, T. T. Liu, L. Liu, G. Chen, J. Ye, G. Cheng, L. B. Cui, H. J. Zhang, W. Qin, and H. Yin. Radiomics signature: A potential biomarker for the prediction of MGMT promoter methylation in glioblastoma. *J Magn Reson Imaging*, 47(5):1380–1387, 05 2018.
- [6] Jiangwei Lao, Yinsheng Chen, Zhi-Cheng Li, Qihua Li, Ji Zhang, Jing Liu, and Guangtao Zhai. A deep learning-based radiomics model for prediction of survival in glioblastoma multiforme. *Scientific Reports*, 7(1):1–8, 2017.
- [7] Zeju Li, Yuanyuan Wang, Jinhua Yu, Yi Guo, and Wei Cao. Deep learning based radiomics (dlr) and its usage in noninvasive idh1 prediction for low grade glioma. *Scientific Reports*, 7(1):1–11, 2017.
- [8] Afshar Parnian, Mohammadi Arash, Pascal N Tyrrell, Patrick Cheung, Sigiuk Ahmed, Konstantinos N Plataniotis, Elsie T Nguyen, and Oikonomou Anastasia. Drtop: deep learning-based radiomics for the time-to-event outcome prediction in lung cancer. *Scientific Reports*, 10(1), 2020.
- [9] Yen-Tsung Huang and Wen-Chi Pan. Hypothesis test of mediation effect in causal mediation model with high-dimensional continuous mediators. *Biometrics*, 72(2):402–413, 2016.
- [10] Max T Aung, Yanyi Song, Kelly K Ferguson, David E Cantonwine, Lixia Zeng, Thomas F McElrath, Subramaniam Pennathur, John D Meeker, and Bhramar Mukherjee. Application of an analytical framework for multivariate mediation analysis of environmental data. *Nature Communications*, 11(1):1–13, 2020.
- [11] Oliver Y Chén, Ciprian Crainiceanu, Elizabeth L Ogburn, Brian S Caffo, Tor D Wager,

- and Martin A Lindquist. High-dimensional multivariate mediation with application to neuroimaging data. *Biostatistics*, 19(2):121–136, 2018.
- [12] Uri Alon. *An introduction to systems biology: design principles of biological circuits*. CRC press, 2019.
- [13] Linda M Collins and Stephanie T Lanza. *Latent class and latent transition analysis: With applications in the social, behavioral, and health sciences*, volume 718. John Wiley & Sons, 2009.
- [14] Karen Bandeen-Roche, Diana L Miglioretti, Scott L Zeger, and Paul J Rathouz. Latent variable regression for multiple discrete outcomes. *Journal of the American Statistical Association*, 92(440):1375–1386, 1997.
- [15] Roderick JA Little and Donald B Rubin. *Statistical analysis with missing data*, volume 793. John Wiley & Sons, 2019.
- [16] Clifford C Clogg. Latent class models. In *Handbook of statistical modeling for the social and behavioral sciences*, pages 311–359. Springer, 1995.
- [17] Megan S Schuler, Jeannie-Marie S Leoutsakos, and Elizabeth A Stuart. Addressing confounding when estimating the effects of latent classes on a distal outcome. *Health Services and Outcomes Research Methodology*, 14(4):232–254, 2014.
- [18] Tihomir Asparouhov and Bengt Muthén. Auxiliary variables in mixture modeling: Three-step approaches using m plus. *Structural Equation Modeling*, 21(3):329–341, 2014.
- [19] Donald B Rubin. Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of Educational Psychology*, 66(5):688, 1974.
- [20] Paul W Holland. Statistics and causal inference. *Journal of the American Statistical Association*, 81(396):945–960, 1986.
- [21] Prasad Naik and Chih-Ling Tsai. Partial least squares estimator for single-index models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 62(4):763–771, 2000.
- [22] Nick Patterson, Alkes L Price, and David Reich. Population structure and eigenanalysis. *PLoS genetics*, 2(12):e190, 2006.

- [23] Michael P Epstein, Andrew S Allen, and Glen A Satten. A simple and improved correction for population stratification in case-control studies. *The American Journal of Human Genetics*, 80(5):921–930, 2007.
- [24] Y Zhang, T Cai, S Yu, et al. Methods for high-throughput phenotyping with electronic medical record data using a common semi-supervised approach (phecap). *Nature Protocols*, 14(12):3426–3444, 2019.
- [25] Hiroko Ohgaki. Epidemiology of brain tumors. *Cancer Epidemiology*, pages 323–342, 2009.
- [26] Joost JM Van Griethuysen, Andriy Fedorov, Chintan Parmar, Ahmed Hosny, Nicole Aucoin, Vivek Narayan, Regina GH Beets-Tan, Jean-Christophe Fillion-Robin, Steve Pieper, and Hugo JWL Aerts. Computational radiomics system to decode the radiographic phenotype. *Cancer research*, 77(21):e104–e107, 2017.
- [27] Lu Zhang, Yinghui Ge, Qiuru Gao, Fei Zhao, Tianming Cheng, Hailiang Li, and Yuwei Xia. Machine learning-based radiomics nomogram with dynamic contrast-enhanced mri of the osteosarcoma for evaluation of efficacy of neoadjuvant chemotherapy. *Frontiers in Oncology*, page 4748.
- [28] Kosuke Imai and David A Van Dyk. Causal inference with general treatment regimes: Generalizing the propensity score. *Journal of the American Statistical Association*, 99(467):854–866, 2004.
- [29] Paul R Rosenbaum and Donald B Rubin. The central role of the propensity score in observational studies for causal effects. *Biometrika*, 70(1):41–55, 1983.
- [30] Jared K Lunceford and Marie Davidian. Stratification and weighting via the propensity score in estimation of causal treatment effects: a comparative study. *Statistics in medicine*, 23(19):2937–2960, 2004.
- [31] Grant Henning. Meanings and implications of the principle of local independence. *Language testing*, 6(1):95–108, 1989.
- [32] William Meredith. Measurement invariance, factor analysis and factorial invariance. *Psychometrika*, 58(4):525–543, 1993.
- [33] Sander Greenland, Judea Pearl, and James M Robins. Causal diagrams for epidemiologic research. *Epidemiology*, pages 37–48, 1999.

- [34] Judea Pearl. *Causality*. Cambridge university press, 2009.
- [35] Kenneth A Bollen and Judea Pearl. Eight myths about causality and structural equation models. In *Handbook of causal analysis for social research*, pages 301–328. Springer, 2013.
- [36] Inge S Helland. On the structure of partial least squares regression. *Communications in statistics-Simulation and Computation*, 17(2):581–607, 1988.
- [37] Inge S Helland. Partial least squares regression and statistical models. *Scandinavian Journal of Statistics*, pages 97–114, 1990.
- [38] Mervyn Stone and Rodney J Brooks. Continuum regression: cross-validated sequentially constructed prediction embracing ordinary least squares, partial least squares and principal components regression. *Journal of the Royal Statistical Society: Series B (Methodological)*, 52(2):237–258, 1990.
- [39] William F Massy. Principal components regression in exploratory statistical research. *Journal of the American Statistical Association*, 60(309):234–256, 1965.
- [40] Bhupinder S Dayal and John F MacGregor. Improved pls algorithms. *Journal of Chemometrics*, 11(1):73–85, 1997.
- [41] Ron Wehrens and B-H Mevik. The pls package: principal component and partial least squares regression in r. 2007.
- [42] Bing Li. *Sufficient Dimension Reduction: Methods and Applications with R*. CRC Press, 2018.
- [43] David R Brillinger. A generalized linear model with “gaussian” regressor variables. In *Selected Works of David Brillinger*, pages 589–606. Springer, 2012.
- [44] Debashis Ghosh. Sufficient dimension reduction: An information-theoretic viewpoint. *Entropy*, 24(2):167, 2022.
- [45] Bradley Efron. Nonparametric estimates of standard error: the jackknife, the bootstrap and other methods. *Biometrika*, 68(3):589–599, 1981.
- [46] Sündüz Keleş and Mark R Segal. Residual-based tree-structured survival analysis. *Statistics in Medicine*, 21(2):313–326, 2002.
- [47] Lenny Dang, Shengfang Jin, and Shinsan M Su. Idh mutations in glioma and acute myeloid leukemia. *Trends in molecular medicine*, 16(9):387–397, 2010.

- [48] Adam L Cohen, Sheri L Holmen, and Howard Colman. Idh1 and idh2 mutations in gliomas. *Current neurology and neuroscience reports*, 13(5):345, 2013.

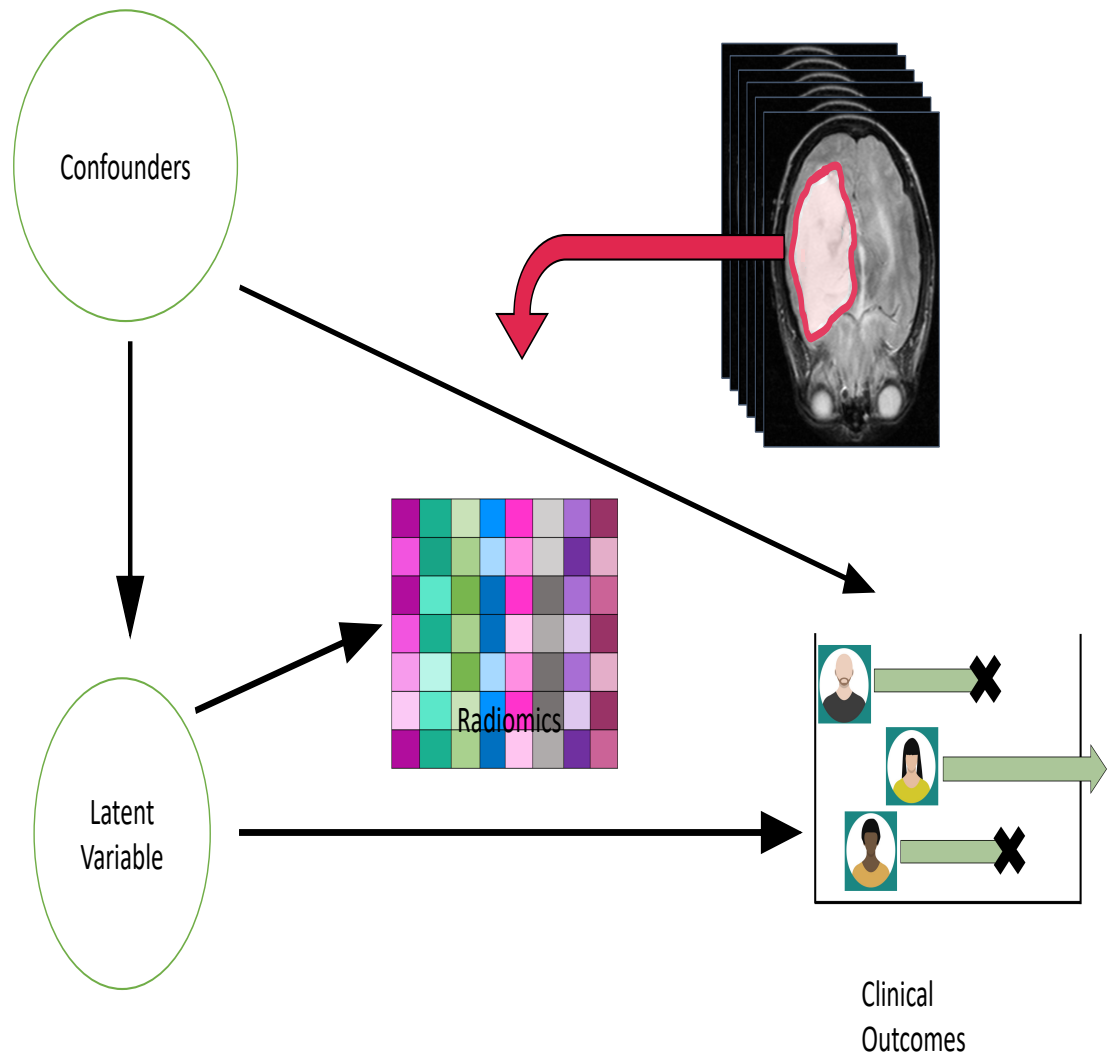


Figure 1: A conceptual model diagram relating confounders, radiomics and outcome variables in medical studies. The goal is to estimate the causal effect corresponding from the arrow from the ‘Latent variable’ circle to the clinical outcome

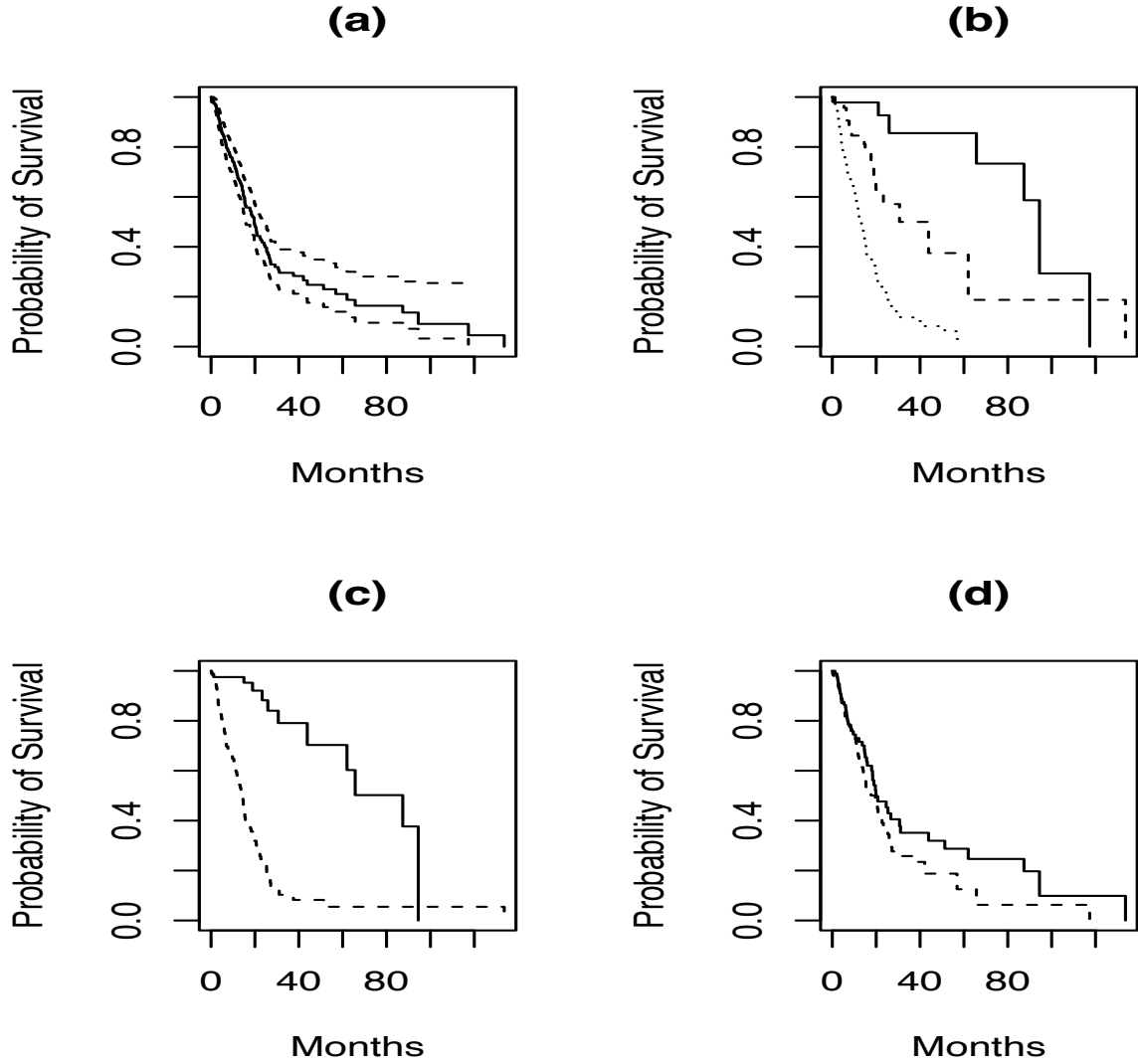


Figure 2: Survival distribution plots for the subjects in the glioblastoma radiomics study. For all plots, the x axis represents the time in weeks and the y-axis is the survival probability. Figure (a) shows the Kaplan-Meier plot for the entire population, along with associated 95% pointwise confidence intervals. Figure (b) shows the Kaplan-Meier estimates by grade (solid = grade 2; dashed = grade 3; dotted = grade 4). In Figure (c), the survival distributions by IDH mutation status (solid = mutant; dashed = wild-type) are presented. Finally, the gender-specific survival distributions (solid=female; dashed = male) are given in Figure d.

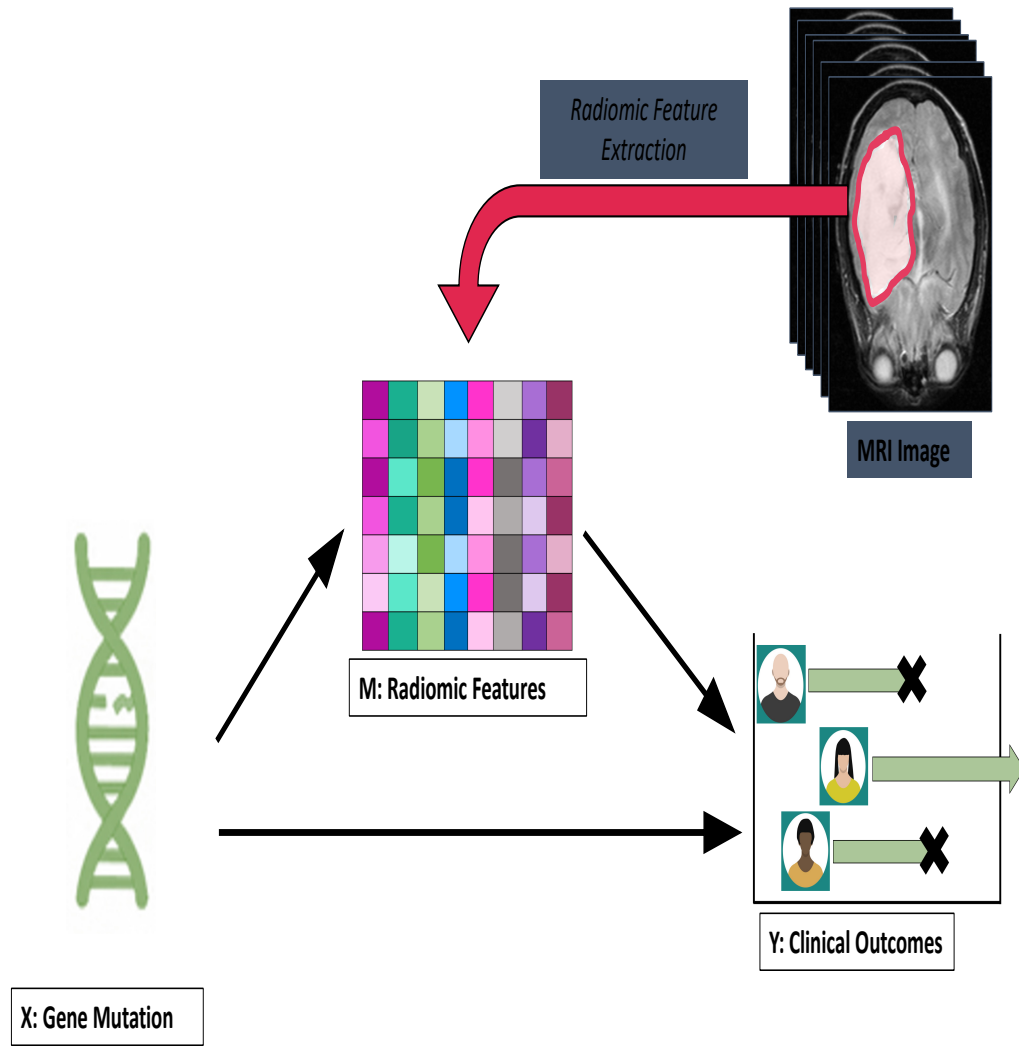


Figure 3: A framework for viewing the radiomics measurements as mediation variables. The exposure here is a DNA mutation which leads to tumorigenesis that is captured by the imaging and radiomics feature and which leads to a clinical outcome.

Table 1: Latent causal effects and associated confidence intervals in glioblastoma study.

Confounders	n	Estimate	95% Confidence Interval
Gender,Grade	225	-0.19	(-0.27,-0.08)
Gender,Grade,IDH	203	-0.20	(-0.26,-0.11)

The estimates denote the average latent local causal effect of radiomics on survival. There are two analyses reported here. The first row denotes an analysis in which gender and grade are confounders. The second row represents an analysis in which gender, grade and IDH mutation status are confounders. The n refers to the sample size used. The estimate is the estimated causal effect using the methods in §4. The 95% confidence intervals are obtained using the nonparametric bootstrap.