Distribution Agreement

In presenting this thesis or dissertation as a partial fulfillment of the requirements for an advanced degree from Emory University, I hereby grant to Emory University and its agents the non-exclusive license to archive, make accessible, and display my thesis or dissertation in whole or in part in all forms of media, now or hereafter known, including display on the world wide web. I understand that I may select some access restrictions as part of the online submission of this thesis or dissertation. I retain all ownership rights to the copyright of the thesis or dissertation. I also retain the right to use in future works (such as articles or books) all or part of this thesis or dissertation.

Signature:

_____ _____

Yi Guo Date

Studying the Association between Overall Survival and Metastatic Sites by Breast Cancer Subtypes Based on National Cancer Database

By

Yi Guo

Master of Science in Public Health

Department of Biostatistics and Bioinformatics

_____

Limin Peng, Ph.D.

Committee Chair

_____

Xiaoxian (Bill) Li, M.D., Ph.D.

Committee Member

Studying the Association between Overall Survival and Metastatic Sites by Breast Cancer Subtypes Based on National Cancer Database

By

Yi Guo

B.S., Sun Yat-sen University, 2015

M.Eng., Cornell University, 2016

Thesis Committee Chair: Limin Peng, Ph.D.

An abstract of

A thesis submitted to the Faculty of the

Rollins School of Public Health of Emory University

in partial fulfillment of the requirements for the degree of

Master of Science in Public Health

in Biostatistics

2019

Abstract

Studying the Association between Overall Survival and Metastatic Sites by Breast Cancer Subtypes Based on National Cancer Database

By Yi Guo

Background: Metastatic breast cancer is the main cause of breast cancer-associated deaths and receptor status has a large impact on prognosis. Other prognostic factors such as adjuvant systemic treatment also contribute to the heterogeneity in breast carcinomas. Within this scenario, how breast cancer subtype and metastatic site are associated with breast cancer overall survival remains unclear.

Method: A total of 5211 female patients with stage IV breast cancer from the National Cancer Database (NCDB) (2010-2013) were examined. All patients received surgery and systemic treatment. The distribution of metastatic sites among breast cancer subtypes was evaluated through a χ^2test. Univariate and multivariate analyses using two semi-parametric approaches, including the Cox proportional hazard (PH) analyses and the censored quantile regression analyses, were conducted to assess the associations between metastatic sites and overall survival.

Results: HR+/HER2- breast cancer was most likely to metastasize to bone, TNBC was most likely to metastasize to brain or lung and HR-/HER2+ was most likely to metastasize to liver or multiple organs. Overall, patients with bone metastasis appeared to have the best prognosis while patients with multiple metastasis had the worst prognosis. In univariate quantile regression analyses, the survival differences between bone metastasis versus multiple metastasis were varied over quantiles, except for HR-/HER2+ breast cancer. In multivariate analysis, age showed negative prognostic effect among patients with all subtypes and was varied in HR+/HER2- subtype. In particular, TNBC patients with bone metastasis versus multiple metastases had varying quantile effects above the median.

Conclusion: This study showed different breast cancer subtypes had different metastatic patterns and survivals. Compared with the Cox model, the censored quantile regression model revealed a more comprehensive prognostic patterns in metastatic breast cancer. Adjusting clinical surveillance and treatment strategies were suggested based on the variation of prognostic effects in different metastatic sites over time.

Studying the Association between Overall Survival and Metastatic Sites by Breast Cancer Subtypes Based on National Cancer Database

By

Yi Guo

B.S., Sun Yat-sen University, 2015

M.Eng., Cornell University, 2016

Thesis Committee Chair: Limin Peng, Ph.D.

A thesis submitted to the Faculty of the
Rollins School of Public Health of Emory University
in partial fulfillment of the requirements for the degree of
Master of Science in Public Health
in Biostatistics
2019

Acknowledgement

Friedrich Nietzsche said, “That which does not kill us makes us stronger.” Life is always challenging and unpredictable. Used to be an engineer, I once thought it was a dream that I can continue studying my favorite subject, math. Encouraged by my grandfather who passed away from cancer, I can even do statistical analyses in cancer data now. Thanks to everyone who helped me selflessly and showed kindness to me during my life and academic studies. My projects and thesis might not have been completed without the support, patience and care from many lovely people.

Firstly, I’m very thankful to my faculty advisor as well as my thesis advisor, Dr. Limin Peng, for giving me insightful academic suggestions throughout the whole academic years, and guiding me during my practicum projects, publication and thesis writings. Although you are busy and told me “you should work hard” during the first year, you are always willing to spare some time to answer my questions related to my thesis and projects. Without your excellent academic guidance and advice for my life, I could not realize more shortcomings from myself and improve them only by working hard. Many thanks to you!

I’m also grateful to my committee member, Dr. Xiaoxian Li, for your patient guidance and encouragement during my practicum. Thank you for your quick response to all of my emails and questions during the busy semesters. You provided me many opportunities to enhance my practical skills even though I was a new man. I was inspired from your enthusiasm, creative thoughts, high efficiency and intuition in your research work, and also learned a lot of knowledge in the medical field. These are of great help for me in the future to do more clinical meaningful analyses in statistics.

I would like to thank all the staffs and faculties in Department of Biostatistics and Bioinformatics. Thank you for providing me a lot of resources and helps in my academic life.

Finally, I appreciate my parents, classmates and friends, who encouraged me whenever I was facing dilemmas in my projects and thesis writing. You always gave me confidence and hope when I felt upset. Special thanks to my parents for completely understanding and supporting any of my personal decision. I love you all.

Table of Contents

LIST OF TABLES

LIST OF FIGURES

CHAPTER I: INTRODUCTION

1.1 Background

Breast cancer is the most common cancer and the leading cause of cancer death among women worldwide.[1] In 2018, there were an estimated 266,120 females (15.3% of all new cancer cases) diagnosed with breast cancer and an estimated 40,920 females died from the disease in the United States.[2] Breast cancer results from the uncontrolled growth of tumor cells. Depending on how far the cells spread, it can be divided into three main stages: localized, regional, distant.[2] In the localized stage, the cells grow out of control but within the breast. In the regional stage, the cells spread outside of the primary site but only metastasize to the regional lymph nodes. While in the distant stage, the spread of breast cancer is beyond the regional lymph node, indicating distant metastasis.[2]

Distant metastasis (Stage IV breast cancer) is caused by the spreading or metastasizing of cells from the breast through the blood or lymphatic system and colonizing in distant parts of the body, most commonly the bone, brain, liver and lung. Stage IV breast cancer is the leading cause of breast cancer-associated deaths.[3] In 2018, the 5-year survival rate of female breast cancer regardless of the cancer stage was estimated to be 89.7% in the U.S. However, the 5-year survival rate at stage IV was estimated to be 27.0%, indicating only 27.0 out of 100 females with metastatic breast cancer is expected to be alive five years after diagnosis.[2] These 5-year survival rates indicate that breast cancers below stage III usually can be successfully treated, whereas cases in stage IV continue to have high mortality rates.[4] To further understand this poorer prognosis associated with stage IV breast cancers, the focus of this study will be on metastatic breast cancer.

Survival outcomes from Stage IV breast cancer are affected by many prognostic factors including the site of metastasis, as well as the receptor status of the tumors.[5-7] These tumor receptors include estrogen receptor (ER), progesterone receptor (PR), and the human epidermal growth factor

receptor 2 (HER2).[8] Either ER+ or PR+ can be defined as HR+. Accordingly, breast cancer can be classified into four main subtypes: HR+/HER2-, HR+/HER2+, HR-/HER2+, HR-/HER2- (also known as triple negative breast cancer, TNBC).[6] These breast cancer subtypes have different biology and show differences in tumor growth.[9] For instance, ER-positive (ER+) cancer has better prognosis while triple negative breast cancer (TNBC) has worse prognosis with poorer overall survival.[10-11] This is largely attributed to TNBC lacking the expression of ER, PR and HER2 receptors, termed HR-/HER2-, which inevitably accelerates the growth of tumors.[12-14]

Meanwhile, many studies examined the correlations between receptor status and metastasis. Among ER+ patients, bone is the site of the most frequent metastasis and the brain is the least common site.[3] Among HER2+ patients, ER+ patients were more likely to have bone metastasis but less likely to have brain, liver, lung and multiple metastases in comparison to the ER- patients.[15] TNBC were more likely to have brain, lung metastasis, especially among female patients.[16, 17]

Although existing researches demonstrate the different metastatic patterns and survivals associated with different breast cancer subtypes, most published studies to date have used either small sample sizes or samples without wide coverage. More importantly, they assumed the difference between prognostic effects were constant over survival time, which could only be rough estimates under the condition of extensive genetic heterogeneity in breast carcinomas.[18, 19] Furthermore, patient's age, staging, adjuvant treatment and other prognostic factors also play an important role in survival prediction.[19] Adjuvant therapy is treatment given in addition to the surgery.[20] It includes systemic therapy such as chemotherapy, hormonal therapy, targeted therapies.[21, 22] Within this heterogeneous scenario, it is often challenging for clinicians to make earlier predictions about whether patients need more aggressive forms of therapy.

This thesis project aims to provide more robust information on the effects of different metastasis types on survival time through examining a large breast cancer dataset from the National Cancer Database (NCDB) based on two different semi-parametric regression models.

1.2 National Cancer Database

National Cancer Database (NCDB), established in 1989, is a clinical oncology database jointly supported by the American College of Surgeons and the American Cancer Society in the United States.[23] It has patient information including demographics, tumor characteristics, site-specific factors, metastasis indicators, treatment records and survival outcomes.[24] Breast cancer data from NCDB are currently available from year 2004 to 2016. These hospital-based data were collected from about 34 million current medical records in a nationwide range of over 1,500 accredited facilities.

Since data are reported by multi-institution registries, NCDB has its own data properties of pooling and national standardization. It reflects more strengths in tracing breast cancer etiology and conducting reliable analyses.[25] In comparison to other population-based registries, such as the Surveillance, Epidemiology, and End Result (SEER) database, NCDB has more extensive coverage of incidence cancers in all types in the U.S.[26-28] It covers more than 70% newly diagnosed patients with cancer every year while SEER covers only 28% of new cancer cases. In addition, NCDB has advantages over SEER on providing more detailed staging, initial treatment and adjuvant therapy information.[29, 30] This more detailed information is useful in reducing sampling errors during our patient selection process. In NCDB, more complete records regarding chemotherapy and radiation therapy are included. Anne-Michelle et al. (2016) showed overall sensitivity of SEER data in either chemotherapy (68%) or radiation therapy (80%) was at the moderate level, and differed in cancer site, stage and patient characteristics.[31] Meanwhile, SEER has no information on HER2 targeted therapy and hormonal therapy. Highlights in data coverage and completeness suggest better representativeness of NCDB in evaluating breast cancer and treatment outcomes at country-level than other registries. Since NCDB is more likely to be comprehensive and representative than any other database, we utilize NCDB for this thesis project.

1.3 Two Semiparametric Regression Models

1.3.1 Cox Proportional Hazards Regression Model

Cox proportional hazard (PH) model, proposed by D.R. Cox (1972), is the most popular regression model for evaluating the effects of risk factors on survival outcomes. The Cox PH model is a semi-parametric model including both parametric and nonparametric components. The standard Cox PH analysis adopts the assumptions of non-informative censoring and proportional hazard.[32]

Under the Cox PH model, the exponentiated regression coefficients are usually interpreted as hazard ratios associated with the corresponding risk factor (or covariate). The baseline hazard function is completely unspecified and represents the hazard function of the survival outcome corresponding the situation with all covariates are set as zeros.[33] In this project, we plan to conduct Cox PH analyses to assess the effects of metastatic sites on the overall survival of breast cancer patients, without or with adjustments for other risk factors such as gender, stage, breast cancer subtype, which are commonly evaluated in medical research. [34-36]

1.3.2 Censored Quantile Regression Model

While the Cox PH model is popular in survival analysis, practitioners may be interested comparing survival differences in the time scale, for example, the restricted mean lifetime.[37] Rather than modeling survival time, Cox model focuses on modeling hazard rate; thus it infers the survival time difference only in an indirect way.[38-39]

Quantile regression offers a flexible tool for assessing the survival time differences. It directly formulates covariate effects on the conditional quantiles of the survival time of interest.[40] Consequently, the regression coefficients have a physical interpretation as covariate effects on the quantiles of the survival time. Besides, quantile regression has advantages in flexibility and robustness.[39, 41] It allows the covariates to have different effects on different quantiles, thereby

accommodating a dynamic association structure between covariates and the survival time, which often manifests the underlying population heterogeneity not captured by the measured covariates.[39]

With randomly censored time-to-event data, there are currently three major approaches within the quantile regression framework: Powell (1986) approach for fixed censoring, Portnoy (2003) approach for random censoring, and Peng and Huang approach (2008) for random censoring.[42] In this study, we adopted the quantile regression method proposed by Peng and Huang (2008) mainly because of its weaker random censoring assumption, and clearly-defined estimation procedure. By gathering quantile-specific covariate effects $\boldsymbol{\beta}(\tau)$, we may acquire a more comprehensive view about the relationship between metastasis type and the overall survival of breast cancer patients.

1.3.3 Motivations for Considering Both Models

The standard Cox PH model captures the covariate effects on the conditional hazard function of the survival time of interest, assuming the constant hazard ratios by covariates over time (i.e. proportional hazards assumption). The proportional hazards assumption can be restricted in practice[43]. Censored quantile regression may serve as a useful alternative to the standard Cox PH analysis when the proportional hazards assumption is not reasonable. The censored quantile regression allows for varying covariate effects, and thus may be less restrictive in some data situations. Censored quantile regression also permits physical interpretations of covariate effects on the time scale, which may be preferable to some practitioners. By considering both Cox PH model and censored quantile regression in this thesis project, we will provide thorough evaluations of the impact of metastasis type on the overall survival of breast cancer patients by breast cancer subtypes from different modeling perspectives and assumptions, thereby delivering more robust implications.

1.4 Outlines of Thesis

The rest of the thesis is organized as follows. In Chapter II we describe the study population, patient selection criteria, data cleaning process and important covariates, as well as statistical methods used for our data analyses. In Chapter III we summarize the results from the statistical analyses separately for the adopted models, the Cox PH model and the censored quantile regression model. In Chapter IV we provide conclusions, along with discussions of potential limitations. We also comment on future work to extend this study.

CHAPTER II: METHOD

In this study, we utilized the data derived from a de-identified NCDB file supported by the Winship Research Informatics Shared Resource of Winship Cancer Institute of Emory University. Patient's demographics, tumor characteristics, site-specific factors, metastasis indicators, treatment received and survival outcomes were included in this de-identified NCDB file.

All statistical analyses were performed using SAS 9.4 software and R Studio (Version 1.1.414; R version 3.3.3 (Another Canoe, (2017)). In particular, survival analysis with quantile regression models was conducted using R package *CQRPH*. *CQRPH* package was specially designed for the method proposed by Peng and Huang (2008). All statistics tests were two-sided. A p-value less than 0.05 was considered as statistically significant.

2.1 Data Set Information

2.1.1 Patient Selection

A total of 5211 de-identified female patients with stage IV breast cancer diagnosed from 2010 to 2013 were selected from the NCDB. In order to avoid heterogeneous treatment effects on survival, we tightened the selection criteria and only chose the patients who received surgery and at least one systemic adjuvant treatment (chemotherapy, immunotherapy, HER2 targeted therapy). All selected patients had their own complete records on age at diagnosis (age), tumor grade, status of HR and HER2 receptor, metastatic site and vital status (death or censoring).

2.1.2 Data Cleaning

To meet the selection criteria, we selected all female patients and removed all the missing or unknown data in regard to prognostic factors we might consider in the statistical analyses. As mentioned above, these variables included age, status of HR and HER2 receptor, site of metastasis and survival outcomes. In addition, targeted patients were not required to have entire information on treatments. We retained the patients with surgery on primary site (breast) and at least one record

of systemic treatment to avoid deleting plethora patients so as to cause unnecessary selection bias. The inclusion and exclusion criteria of patient selection were presented as below (Table 2.1).

Table 2. 1 A summary of patient selection criteria

Inclusion criteria	Exclusion criteria
• Female breast cancer patient • First diagnosis with breast cancer between 2010-2014 • History of systemic therapies (chemotherapy, immunotherapy, HER2-targeted therapy)	• No metastasis to bone, brain, liver and lung • Incomplete data of metastatic sites among patients with only one metastasis record • No surgery taken on breast

2.1.3 Variable Description

For descriptive purposes, patient age was further defined as a binary variable defined as age groups, less than 50 years (< 50 years) and greater than or equal to 50 years (≥ 50 years). Tumor grade was reclassified into three levels (I, II, III) and unknown grade, where we combined poorly differentiated and undifferentiated or anaplastic tumor as Grade III. All patients had stage IV breast cancer. The stage was defined on the basis of the current standard of American Joint Committee on Cancer (AJCC) tumor-node-metastasis (TNM) staging system. Since the criteria from AJCC TMN staging system included the assessments of tumor grades, the tumor grade from the dataset was only summarized for descriptive purposes.

Both HR receptor and HER2 receptor had positive and negative types. For HR receptor, the status depended on the both ER and PR receptors, in which either ER+ or PR + can be defined as HR+. Consequently, the patients were classified into four breast cancer subtypes: HR-/HER2+, HR+/HER2+, HR+/HER2-, HR-/HER2- (also known as triple negative breast cancer, TNBC). According to the number of metastatic sites, we first divided the metastasis into three levels, i.e.,

no metastasis, single metastasis, multiple metastases. We then classified the single metastasis into four detailed categories: bone, brain, liver and lung oligometastasis.

2.2 Statistical Analyses

2.2.1 Descriptive Analysis

Descriptive statistics were reported within and beyond breast cancer subtype (HR+/HER2-, HR+/HER2+, HR-/HER2+, TNBC). Categorical variables were assessed by χ^2 tests and summarized by count and proportion. Continuous variables were evaluated by two-sample t tests and summarized by mean and standard deviation. Demographic and clinicopathologic characteristics in the descriptive table contained age, stage, status of ER, PR, HR, HER2, tumor grade, systemic therapies (chemotherapy, HER2 targeted therapy, immunotherapy), radiation and site of metastases. In particular, a χ^2 test was conducted to evaluate the association between metastatic sites and breast cancer subtypes. Within each subtype, Kaplan-Meier curves were generated to summarize the overall survival (OS) times in patients with different metastatic sites and log rank tests were implemented to measure the differences between or among different Kaplan-Meier curves.

2.2.2 Analyses Based on Cox Proportional Hazard Model

A standard form of Cox proportional hazard (PH) model can be written as the product of a non-parametric baseline hazard $h_0(t)$ and a parametric linear regression of covariates X_i in an exponential function.[32] With real data, a covariate, X_i, which can be either continuous or categorical, may denote prognostic factors, such as patient age at diagnosis, and site of metastasis.

$$h(t|\boldsymbol{X}) = h_0(t)\, \mathrm{e}^{(\beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_p X_p)}, \quad (1)$$

where $\beta_1, \beta_2, \ldots, \beta_p$ are the constant coefficients corresponding to the covariates $X_1, X_2, \ldots, X_p$ respectively.

In this study, we conducted the analyses stratified by breast cancer subtypes (HR+/HER2-, HR+/HER2+, HR-/HER2+, TNBC). Within each subtype, the associations between metastatic site and overall survival were examined. Patient's age at diagnosis (age) was adjusted in multivariate analyses. The prognosis impact of a certain covariate can be assessed and interpreted through the hazard ratio (HR) (e.g. $e^{\beta_i}, i = 1, 2, \dots, p$). For instance, the HR between any of the two metastatic sites can be obtained by setting different reference metastasis categories in Cox PH models. In either univariate or multivariate analysis, a higher risk of the event of interest (e.g. death) is shown when HR is estimated to be significantly higher than 1 and vice versa.

Because of the tied observations from the survival data and random censoring assumption from the Cox PH model, the regression coefficients $\beta_1, \beta_2, \dots, \beta_p$ can be estimated by maximum partial likelihood (MPLE).[44, 45] In this study, we used Breslow's approximation (default in SAS *Proc Phreg*) to obtain MPLEs of $\boldsymbol{\beta} = (\beta_1, \beta_2, \dots, \beta_p)^T$. Given n random observations, let $t_1 < t_2 < \cdots < t_j$ denote the j distinct failure times and d_j denote the number of failures (deaths) at t_j. For model (1), the partial likelihood (PL) applying Breslow's method is given by [45-47]

$$L(\boldsymbol{\beta}) = \prod_{j=1}^{J} L_j(\boldsymbol{\beta}) \approx \prod_{j=1}^{J} \frac{\exp(\boldsymbol{\beta} \sum_{i \in D_j} x_i)}{\left[\sum_{l \in R_j} \exp(\boldsymbol{\beta} x_l)\right]^{d_j}}, \quad (2)$$

where R_j is the risk set at time t_j and D_j is the failure set at time t_j. The hypothesis testing of $H_0: \boldsymbol{\beta} = 0$ can be conducted through Wald, likelihood ratio (LR) or score test. In this case, we performed Wald test (default in SAS *Proc Phreg*) for the hypothesis testing.

(1) Univariate Analysis

Univariate Cox PH regression was conducted separately in each breast cancer subtype. As described in *Variable Description* section, the covariate related to the metastasis was reclassified into five categories: bone, brain, liver, lung, multiple. Assume (X_1, X_2, X_3, X_4) are the dummy variables related to the metastatic sites where $X_1 = 1$ if bone, $X_2 = 1$ if brain, $X_3 = 1$ if liver, $X_4 = 1$ if lung metastasis, and multiple metastases is the reference group where $(X_1, X_2, X_3, X_4) = (0, 0, 0, 0)$. Then a univariate Cox model within a certain breast cancer subtype was conducted as follows:

$$h(t|\boldsymbol{X}) = h_0(t)\, e^{(\beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_4)}, \quad (3)$$

Survival difference compared with multiple metastases (reference group) can be assessed and interpreted through hazard ratio e^{β_i}, where $i = 1, 2, 3, 4$. By defining different metastatic sites as reference category, the HR between each pair of metastatic sites in four subtypes can be obtained based on model (3).

(2) Multivariate Analysis

In the previous section, univariate Cox models were conducted in different breast cancer subtypes to investigate the overall survival (OS) with respect to the site of metastasis. However, it is still of interest to consider the prognostic effects of metastatic sites while adjusting for the impacts of other prognostic factors. Based on the availability of the data, we adjusted for age in the multivariate analysis. Including the age as a prognostic factor, a multivariate Cox regression model based on model (3) can be written as follows:

$$h(t|\boldsymbol{X}) = h_0(t)\, e^{(\beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_4 + \beta_5 X_5)}, \quad (4)$$

where age is represented as a continuous variable X_5.

Similar to the univariate analysis, the HR between any two of the metastatic sites can be calculated by changing the reference categories based on model (4).

2.2.3 Analyses Based on Censored Quantile Regression Model

Conditional quantile is the key concept for understanding quantile regression. For an outcome of interest, denoted by Y, the τth conditional quantile of Y given $\boldsymbol{X}$ is defined as

$$Q_Y(\tau|\boldsymbol{X}) = \inf\{y: P\{Y \leq y \mid X = x\} = \tau\}, \quad (5)$$

where $\tau \in (0,1)$ and $\boldsymbol{X}$ is a known $p \times 1$ covariate vector.

A typical censored quantile regression model assumes that

$$Q_Y(\tau|X) = \boldsymbol{X}^T\boldsymbol{\beta}(\tau), \qquad \tau \in (0,1), \quad (6)$$

with $Y = \log T$, representing the log-transformed survival time T. It can be shown that the accelerated failure time (AFT) model is a special case of model (6).[48] Unlike the Cox PH model, the AFT model shares the same spirit as the classic linear regression and provides more straightforward interpretation in survival time.[39] By allowing the coefficients of $\boldsymbol{X}$, i.e., $\boldsymbol{\beta}(\tau)$, to vary with τ, model (6) can accommodate covariate effects beyond the simple location-shift effects that are assumed under the AFT model, and therefore is more flexible and robust.[41]

Compared with the Cox model, the censored quantile regression model (6) provides an alternative view to assess the impact of covariates on survival outcomes. More specifically, the Cox model depicts how covariates influence the conditional hazard function of T given X, while censored quantile regression implies how covariates change the conditional quantiles of T given X. The Cox PH model imposes location-shift effects of covariates on the log hazard function. In contrast, censored quantile regression permits a more flexible relationship between covariates and quantiles of T.

As mentioned in the Chapter I, we used the quantile regression method proposed by Peng and Huang (2008) to estimate model (6) with randomly censored survival data. In the estimation procedure, Peng and Huang (2008) developed a martingale-based estimating equation extended from the Nelson-Aalen estimator of the cumulative hazard function, which simplifies the algorithms by only minimizing L_1- type convex functions.[42, 49] According to the properties of martingale in survival data, they started the estimation from $E\{M_i(t)|\,\boldsymbol{X}_i\} = 0$ for $t \geq 0$.[49] Here, $M_i(t)$ is the martingale process related to the counting process $N_i(t)$, where $M_i(t) = N_i(t) - \Lambda_T(t \wedge (T_i \wedge C_i)|\,X_i)$ and $\Lambda_T(\cdot\,|\,X_i)$ is the cumulative hazard function of T_i given X_i; $N_i(t) = I(\{(T_i \wedge C_i) \leq t\},\ \{\delta_i = 1\})$ and $\delta_i = I\{T_i < C_i\}$. Due to the stochastic nature of martingale, the $\boldsymbol{\beta}(\tau)$ can be estimated subsequently with a grid-based method. Moreover, the property of monotonicity associated with Peng and Huang (2008)'s estimating equation reduces the algorithmic complexity in solving $\boldsymbol{\beta}(\tau)$ through minimization of L_1- type convex function at a fine, pre-specified τ -grid.[49] Peng and Huang (2008) also established asymptotic properties and developed a resampling approach from Jin et al. (2001) for the inference procedure.[41, 49-50] The resampling-based inference contains: 1) overall significance test of covariate effect over $\tau \in [\tau_L, \tau_U]$, i.e., $H_0: \beta_0^{(i)}(\tau) = 0$, where $0 < \tau_L \leq \tau_U < 1$ and $i = \{2, \ldots, p\}$; 2) constancy test of covariate effect from secondary inference over $\tau \in [\tau_L, \tau_U]$, i.e., $H_0: \beta_0^{(i)}(\tau) = \eta_0$, where η_0 is an unknown constant, $0 < \tau_L \leq \tau_U < 1, i = \{2, \ldots, p\}$; 3) model diagnostics.[49]

(1) Univariate Analysis

With the same formulation of X_i ($i = 1, 2, 3, 4$) as in the univariate Cox analysis, we first fit censored quantile regression model (6) that only adjusted for metastatic sites, i.e.

$$Q_Y(\tau|\boldsymbol{X}) = \beta_0(\tau) + \beta_1(\tau)X_1 + \beta_2(\tau)X_2 + \beta_3(\tau)X_3 + \beta_4(\tau)X_4, \qquad \tau \in (0,1). \quad (7)$$

Or equivalently,

$$Q_T(\tau|\boldsymbol{X}) = e^{\beta_0(\tau)+\beta_1(\tau)X_1+\beta_2(\tau)X_2+\beta_3(\tau)X_3+\beta_4(\tau)X_4}, \qquad \tau \in (0,1). \quad (8)$$

According to Model (7), survival difference (after log transformation) between single metastasis and multiple metastases over τ can be assessed through $\beta_i(\tau)$, where $i = 1, 2, 3, 4$.

(2) Multivariate Analysis

We next fit censored quantile regression model which included age in addition to metastatic sites within each breast cancer subtype. The model can be written below:

$$Q_Y(\tau|\boldsymbol{X}) = \beta_0(\tau) + \beta_1(\tau)X_1 + \beta_2(\tau)X_2 + \beta_3(\tau)X_3 + \beta_4(\tau)X_4 + \beta_5(\tau)X_5, \qquad \tau \in (0,1). \quad (9)$$

Or equivalently,

$$Q_T(\tau|\boldsymbol{X}) = e^{\beta_0(\tau)+\beta_1(\tau)X_1+\beta_2(\tau)X_2+\beta_3(\tau)X_3+\beta_4(\tau)X_4+\beta_5(\tau)X_5}, \qquad \tau \in (0,1). \quad (10)$$

(3) Hypothesis Testing and Second Stage Inference

We also conducted hypothesis testing as well as secondary inferences in quantile regression analyses to examine the variations of coefficient effects over quantile levels. These procedures were of great help in further exploring the internal metastatic patterns associated with survival. The average covariate effects were assessed by overall significance tests and the variation of covariate effects were extrapolated by constancy tests.[50] In constancy tests, we selected weight function 1 as the indicator function provided in *CQRPH* R package, where τ equals to 1 when it was larger or equal than the middle of τ-range.

CHAPTER III: RESULTS

3.1 Descriptive Summary

3.1.1 Demographic and Clinicopathological Characteristics

A total of 5211 female patients with Stage IV breast cancer were considered in this study. The maximum follow-up time was 71.23 months with a median of 29.57 months. Table 3.1 summarized the descriptive statistics including patient age at diagnosis, breast cancer subtype, metastatic site, information of systemic adjuvant treatments. Of these patients, most of them were over 50 years old when they were first diagnosed with breast cancer (71.25%) (Table 3.1). The majority of the patients had HR+/HER2- breast cancer subtypes (3269, 62.73%) while only 459 patients (8.81%) had HR-/HER2+ breast cancer subtypes. Besides, there were 620 patients who had HR+/HER2+ subtype (11.90%) and 863 patients who had TNBC subtype (Table 3.1).

In Table 3.1, distant metastasis happened the most in bone organ (2803, 53.79%), followed in descending order by multiple (1046, 20.07%), lung (735, 14.10%), liver (544, 10.44%) and brain organ (83, 1.59%). All patients had surgery and at least one systemic treatment (chemotherapy, hormonal therapy, HER2 targeted therapy). Regardless of the missing records from these patients, 67.68% received chemotherapy, 74.44% received hormonal therapy, 6.24% received HER2 targeted therapy, and 48.32% received radiation therapy (Table 3.1).

Table 3. 1 Descriptive and clinical characteristics for NCDB breast cancer patients diagnosed from 2010 to 2014

Characteristic	Number of Patients (%) or Mean (±SD)
Age at Diagnosis	58.47±13.30
≤ 50	1498 (28.75)
> 50	3713 (71.25)
Subtype	
HR+/HER2-	3269 (62.73)
HR+/HER2+	620 (11.90)
HR-/HER2+	459 (8.81)
TNBC	863 (16.56)
Metastatic Site	
Bone	2803 (53.79)
Brain	83 (1.59)
Liver	544 (10.44)
Lung	735 (14.10)
Multiple	1046 (20.07)
Grade	
I	358 (6.87)
II	1889 (36.25)
III	2520 (48.36)
Unknown	444 (8.52)
Chemotherapy	
Yes	3527 (67.68)
No	1337 (25.66)
Missing	347 (6.66)
Radiation Therapy	
Yes	2518 (48.32)

No	2650 (50.85)
Missing	43 (0.83)
Hormonal Therapy	
Yes	3879 (74.44)
No	1192 (22.87)
Missing	140 (2.69)
HER2 Targeted Therapy	
Yes	325 (6.24)
No	4871 (93.48)
Missing	15 (0.29)

1. Patients might have received more than one therapy.

3.1.2 Metastatic Patterns in Breast Cancer Subtypes

Table 3.2 showed the different distributions of metastatic sites in breast cancer subtypes. Significant association was shown between site of metastasis and breast cancer subtype ($p < .0001$; Table 3.2). Except for HR-/HER2+ subtype, bone was the most frequent metastatic sites among all subtypes. Patients with HR-/HER2+ subtype had the most frequent metastasis to multiple organs (26.58%).

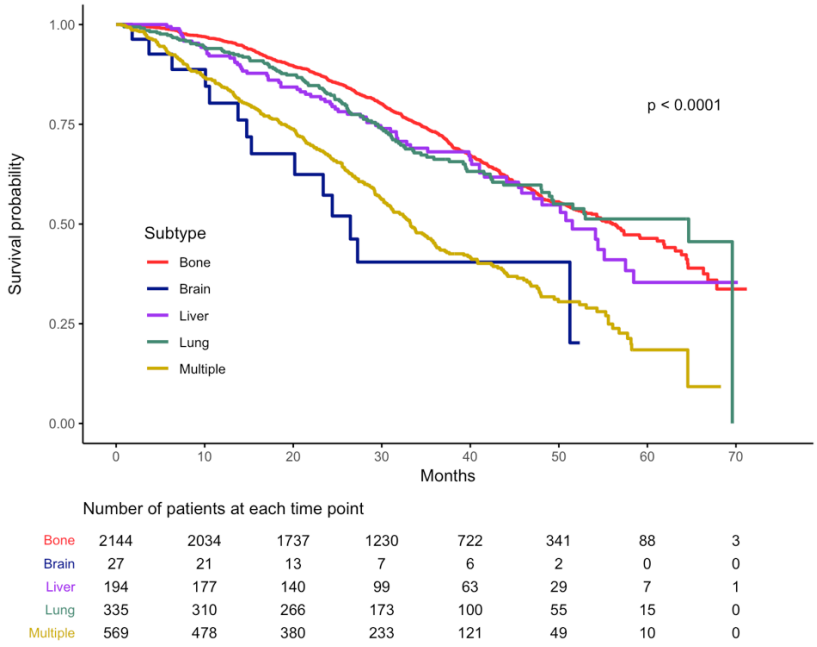
Table 3. 2 Distribution of metastatic sites in breast cancer subtypes

Subtype	Metastatic Site				
	Single Metastasis				Multiple Metastases
	Bone	Brain	Liver	Lung	
HR+/HER2- (n = 3269)	2144 (65.59)	27 (0.83)	194 (5.93)	335 (10.25)	569 (17.41)
HR+/HER2+ (n = 620)	304 (49.03)	7 (1.13)	113 (18.23)	63 (10.16)	133 (21.45)
HR-/HER2+ (n = 459)	112 (24.40)	14 (3.05)	131 (28.54)	80 (17.43)	122 (26.58)
TNBC (n = 863)	243 (28.16)	35 (4.06)	106 (12.28)	257 (29.78)	222 (25.72)

1. The numbers were expressed as absolute number (%).

2. $P < .0001$ in χ^2 test.

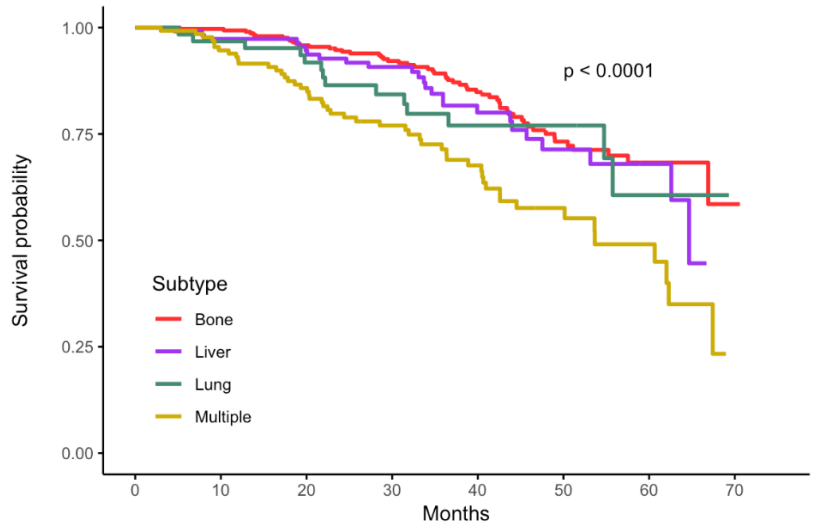
Besides, Kaplan-Meier curves with different metastatic sites were drawn in HR+/HER2-, HR+/HER2+, HR-/HER2+, TNBC subtypes (Figure 3.1). Log rank test showed significant survival difference among metastatic sites within each breast cancer subtype (all $P < .0001$; Figure 3.1).



Number of patients at each time point

Bone	2144	2034	1737	1230	722	341	88	3
Brain	27	21	13	7	6	2	0	0
Liver	194	177	140	99	63	29	7	1
Lung	335	310	266	173	100	55	15	0
Multiple	569	478	380	233	121	49	10	0

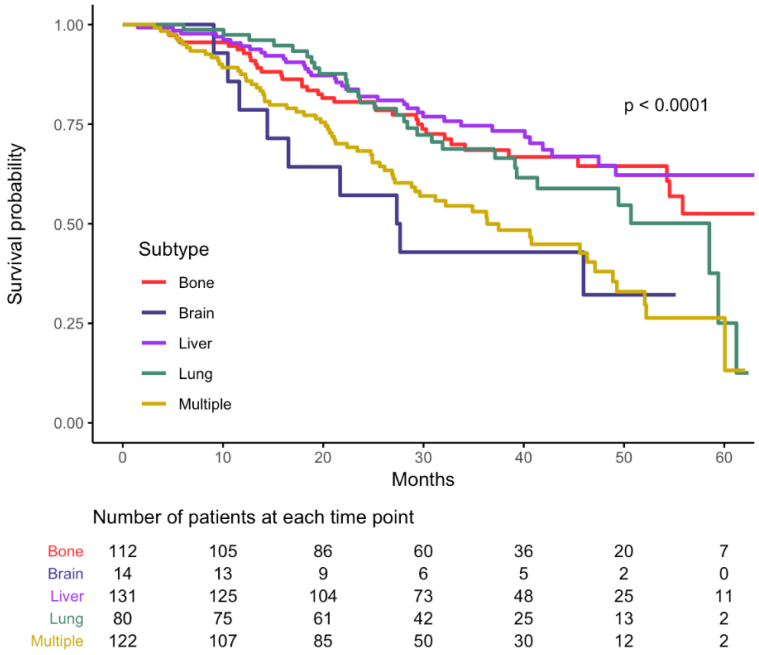
(A) HR+/HER2-



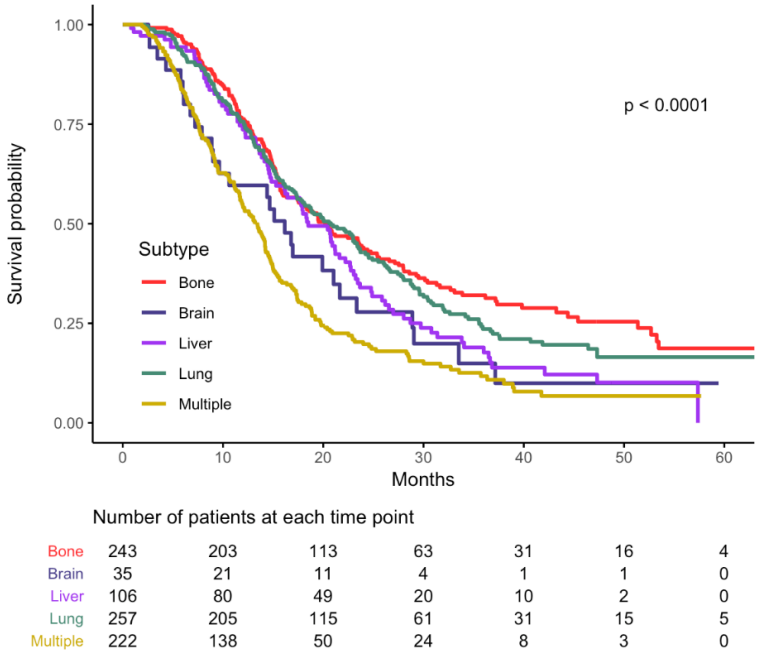
Number of patients at each time point

Bone	304	296	264	203	146	78	28	2
Liver	113	110	102	83	49	25	9	0
Lung	63	60	53	39	24	18	6	0
Multiple	133	123	102	76	50	26	12	0

(B) HR+/HER2+



(C) HR-/HER2+



(D) TNBC

Figure 3. 1 Kaplan-Meier curves of overall survival comparing patients with different metastasis in HR+/HER2- (A), HR+/HER2+ (B), HR-/HER2+ (C) and TNBC (D)

3.2 Cox Proportional Hazard Regression

3.2.1 Univariate Analysis

Breast cancer patients with metastasis to different organs showed different overall survival (Table 3.3). The metastatic patterns associated with overall survival (OS) were similar among all patients except for those with TNBC subtypes.

In HR+/HER2- subtype, patients with bone, liver or lung metastasis had better prognosis than brain metastasis or multiple metastases (Table 3.3). No significant OS differences were found between brain metastasis and multiple metastases, or among bone, liver and lung metastasis (Table 3.3).

In HR+/HER2+ subtype, patients with multiple metastases had worse OS than those with bone, liver or lung metastasis (Table 3.3). Similar as HR+/HER2- subtype, bone, liver and lung metastasis had no significant difference in OS. However, all comparisons with brain metastasis were neglected in HR+/HER2+ subtype due to small sample size (n = 7).

HR-/HER2+ subtype had similar metastatic patterns as HR+/HER2-. The OS with multiple metastases or brain metastasis was significantly worse than that with metastasis to bone, liver or lung (Table 3.3). Similarly, the comparison in OS between neither brain metastasis and multiple metastases, nor bone, liver and lung metastasis was non-significant.

Unlike other subtypes, TNBC subtype had different metastatic patterns associated with OS. Patients with bone metastasis had significantly longer OS than those with lung, brain or multiple metastases; while multiple metastases had the significantly lower OS than bone, liver or lung metastasis (Table 3.3). Bone tended to have the longest survival among all metastatic sites (bone vs lung: HR = 0.9, P = .309; Table 3.3) and multiple metastatic site appeared to have the shortest OS (brain vs multiple: HR = 0.79, P = .237; Table 3.3).

3.2.2 Multivariate Analysis

In the previous section, univariate Cox model within each subtype was conducted to investigate the overall survival (OS) with respect to the site of metastasis. However, it is still necessary to consider the impacts of other prognostic factors at the same time. Due to the strict inclusion and exclusion criteria for the study patients (Table 2.1), we only adjusted for age in the multivariate analyses.

In multivariate analyses adjusting for age, the metastatic patterns within HR+/HER2- and TNBC subtype were not totally the same as the results from univariate analyses (Table 3.4). All significant comparison results from both univariate and multivariate analyses were summarized in Table 3.5.

In HR+/HER2- subtype, patients with bone metastasis had the best OS, followed by patients with lung, liver, brain, multiple metastases in a descending order (Table 3.4). Particularly, patients with brain metastasis and multiple metastases had no difference in OS and showed equally worse outcome compared with other metastatic sites (Brain vs Multiple: HR = 1.24, p = .241; Table 3.4).

The OS for HR+/HER2+ and HR-/HER2+ metastatic breast cancer patients in multivariate analysis had the same metastatic patterns as the results in univariate analysis (Table 3.4, Table 3.5). Either in HR+/HER2+ or HR-/HER2+ subtype, breast cancer metastasizing in bone, liver or lung organ had better OS in multiple organs (or brain).

Among patients with TNBC subtype, bone metastasis had the better OS than any other metastasis (Table 3.4). Overall survival in patients with multiple metastases was significantly shorter than bone, liver or lung oligometastasis.

In conclusion, breast cancer patients who had surgery and at least one systemic therapy (chemotherapy, hormone therapy, HER2 targeted therapy) in all subtypes tended to have better prognosis when they only had single bone metastasis, and worse prognosis when they had single brain metastasis or multiple metastases.

Table 3. 3 Univariate analysis for overall survival comparing sites of metastasis in different subtypes using Cox PH models

Metastatic Site	HR+/HER2-		HR+/HER2+		HR-/HER2+		TNBC	
	HR (95% CI)	P	HR (95% CI)	P	HR (95% CI)	P	HR (95% CI)	P
Bone vs. Brain	0.35 (0.21-0.60)	.0001*	1.54 (0.21-11.16)	.667	0.43 (0.21-0.90)	.024*	0.62 (0.41-0.92)	.018*
Bone vs. Liver	0.82 (0.64-1.05)	.110	0.80 (0.50-1.28)	.356	1.23 (0.77-1.95)	.382	0.74 (0.57-0.96)	.022*
Bone vs. Lung	0.91 (0.75-1.11)	.359	0.78 (0.43-1.40)	.409	0.82 (0.51-1.34)	.435	0.90 (0.73-1.11)	.309
Bone vs. Multiple	0.44 (0.38-0.50)	<.0001*	0.41 (0.28-0.60)	<.0001*	0.50 (0.33-0.75)	.0008*	0.48 (0.39-0.60)	<.0001*
Brain vs. Liver	2.32 (1.31-4.12)	.004*	0.52 (0.07-3.83)	.519	2.85 (1.37-5.92)	.005*	1.20 (0.78-1.84)	.407
Brain vs. Lung	2.58 (1.48-4.50)	.0008*	0.51 (0.07-3.85)	.510	1.91 (0.91-4.02)	.089*	1.46 (0.98-2.17)	.063
Brain vs. Multiple	1.23 (0.72-2.10)	.452	0.26 (0.04-1.92)	.188	1.15 (0.57-2.31)	.695	0.79 (0.53-1.17)	.237
Liver vs. Lung	1.11 (0.83-1.50)	.481	0.98 (0.51-1.88)	.942	0.67 (0.41-1.09)	.107	1.22 (0.94-1.57)	.133
Liver vs. Multiple	0.53 (0.41-0.69)	<.0001*	0.52 (0.32-0.83)	.006*	0.41 (0.27-0.61)	<.0001*	0.66 (0.51-0.85)	.001*
Lung vs. Multiple	0.48 (0.39-0.59)	<.0001*	0.52 (0.29-0.95)	.032*	0.60 (0.39-0.93)	.022*	0.54 (0.44-0.66)	<.0001*

1. * Significant P value (<0.05).

2. Abbreviation: HR, hazard ratio.

Table 3. 4 Multivariate analysis for overall survival comparing sites of metastasis in different subtypes using Cox PH models

Metastatic Site	HR+/HER2-		HR+/HER2+		HR-/HER2+		TNBC	
	HR (95% CI)	P	HR (95% CI)	P	HR (95% CI)	P	HR (95% CI)	P
Age	1.03 (1.02-1.03)	<.0001*	1.04 (1.03-1.05)	<.0001*	1.03 (1.01-1.04)	.0005*	1.01 (1.00-1.01)	<.0001*
Bone vs. Brain	0.36 (0.21-0.61)	.0002*	1.55 (0.21-11.21)	.664	0.41 (0.20-0.86)	.017*	0.62 (0.41-0.92)	.019*
Bone vs. Liver	0.74 (0.58-0.94)	.014*	0.72 (0.45-1.15)	.167	1.17 (0.73-1.85)	.516	0.74 (0.56-0.94)	.016*
Bone vs. Lung	1.03 (0.84-1.26)	.779	0.85 (0.48-1.54)	.598	0.87 (0.53-1.41)	.566	0.90 (0.73-1.11)	.335
Bone vs. Multiple	0.44 (0.38-0.51)	<.0001*	0.42 (0.28-0.61)	<.0001*	0.48 (0.32-0.73)	.0005*	0.48 (0.39-0.59)	<.0001*
Brain vs. Liver	2.04 (1.15-3.61)	.015*	0.46 (0.06-3.41)	.449	2.83 (1.36-5.90)	.005*	1.17 (0.77-1.80)	.466
Brain vs. Lung	2.85 (1.64-4.96)	.0002*	0.55 (0.07-4.20)	.565	2.11 (1.00-4.46)	.051	1.46 (0.98-2.17)	.063
Brain vs. Multiple	1.23 (0.72-2.10)	.451	0.27 (0.04-1.96)	.196	1.18 (0.58-2.36)	.651	0.78 (0.52-1.15)	.209
Liver vs. Lung	1.40 (1.04-1.89)	.027*	1.19 (0.62-2.30)	.600	0.74 (0.46-1.21)	.237	1.25 (0.96-1.61)	.094
Liver vs. Multiple	0.60 (0.47-0.78)	.0001*	0.59 (0.36-0.95)	.030*	0.42 (0.28-0.63)	<.0001*	0.66 (0.51-0.85)	.002*
Lung vs. Multiple	0.43 (0.35-0.54)	<.0001*	0.49 (0.27-0.89)	.018*	0.56 (0.36-0.86)	.009*	0.53 (0.43-0.65)	<.0001*

1. * Significant P value (<0.05).

2. Abbreviation: HR, hazard ratio.

Table 3. 5 Overall survival comparisons between metastatic sites from both univariate and multivariate analysis using Cox PH models

Subtype	Univariate Analysis	Multivariate Analysis
HR+/HER2-	(Bone, Liver, Lung) > (Brain, Multiple)	Bone > Lung > Liver > (Brain, Multiple)
HR+/HER2+	(Bone, Liver, Lung) > Multiple	Same as univariate results
HR-/HER2+	(Bone, Liver, Lung) > (Brain, Multiple)	Same as univariate results
TNBC	Bone > (Liver, Brain); (Lung, Brain) > Multiple	Bone > (Liver, Lung) > Multiple; Bone > Brain

1. Metastatic sites within the same parenthesis had non-significant OS difference between each other ($P > 0.05$).

3.3 Censored Quantile Regression

In this section, we fit our dataset with censored quantile regression models using Peng and Huang (2008)'s method. Hypothesis testing and second stage inference were also performed to have further exploration in varying quantile effects of covariates. Since few events happened around time at 0, causing unstable and biased estimates, we only consider the situation where $\tau > 0.1$ in all the following analyses.

3.3.1 Univariate Analysis

Model (8) was performed within each breast cancer subtype (HR+/HER2-, HR+/HER2+, HR-/HER2+, TNBC) to compare the survival time between patients with single site of metastasis (bone, brain, liver, lung) versus the reference group (multiple metastatic sites). Average covariate effects, overall significance, and constancy of covariate effects over τ in different breast cancer subtypes were shown in Table 3.6. Figure 3.2-3.3 displayed the coefficient estimates (multiple metastases as reference group) with 95% pointwise confident intervals over quantile level τ. All subplots with different reference categories of metastatic sites were attached in the Appendix A.

Table 3. 6 Estimates of average covariate effects and results on hypothesis testing and second-stage inference in univariate quantile regression model stratified by breast cancer subtypes

	HR+/HER2-		
Metastatic Site	**Average Effect** $\tau\in[0.1,0.6]$	**Overall Significance** $\tau\in[0.1,0.6]$	**Constancy** $\tau\in[0.1,0.6]$
Bone	0.583	< 0.0001*	0.003*
Brain	-0.170	0.461	0.988
Liver	0.465	< 0.0001*	0.707
Lung	0.528	< 0.0001*	0.293
	HR+/HER2+		
Metastatic Site	**Average Effect** $\tau\in[0.1,0.35]$	**Overall Significance** $\tau\in[0.1,0.35]$	**Constancy** $\tau\in[0.1,0.3]$
Bone	0.538	0.0002*	0.115
Liver	0.370	0.093	0.790
Lung	0.459	0.003*	0.212
	HR-/HER2+		
Metastatic Site	**Average Effect** $\tau\in[0.1,0.35]$	**Overall Significance** $\tau\in[0.1,0.35]$	**Constancy** $\tau\in[0.1,0.35]$
Bone	0.398	0.050*	0.485
Brain	-0.133	0.383	0.397
Liver	0.547	0.002*	0.639
Lung	0.457	0.005*	0.334
	TNBC		
Metastatic Site	**Average Effect** $\tau\in[0.1,0.75]$	**Overall Significance** $\tau\in[0.1,0.75]$	**Constancy** $\tau\in[0.1,0.75]$
Bone	0.524	< 0.0001*	0.284
Brain	0.127	0.414	0.164
Liver	0.387	< 0.0001*	0.740
Lung	0.473	< 0.0001*	0.371

1. * Significant P value (<0.05).

2. Reference category: multiple metastases.

3. Average covariate effects were resampling based.

(1) HR+/HER2- Subtype

The estimates and 95% confidence intervals of proportional survival time after logarithm between single sites of metastases and multiple sites of metastases for $\tau \in [0.1, 0.639]$ were visualized in Figure 3.2.

Compared to the multiple metastases, bone, liver or lung metastasis had significantly better survival over $\tau \in [0.1, 0.6]$ (all P < .0001; Table 3.6). Besides, there is little difference between brain metastasis and multiple metastases (P = 0.461; Table 3.6) in terms of survival quantiles.

Constancy tests over $\tau \in [0.1, 0.6]$ showed the covariate effect related to bone may be different across different quantile levels (P =.003; Table 3.6). It decreased over quantiles in Figure 3.2, meaning the survival difference between bone metastasis and multiple metastases became smaller among long-term survivors.

In addition, the difference in prognosis between patients with lung metastasis and multiple metastases were constant over $\tau \in [0.1, 0.6]$ (P = 0.293; Table 3.6). However, the fluctuation of the pattern related to the lung and multiple metastases was visualized in Figure 3.2. After splitting the τ range into two smaller intervals, i.e., $[0.1, 0.3]$ and $[0.3, 0.55]$, we conducted two constancy tests separately and found significant variations within both quantile intervals (P = .004; P = .007). The significant constant effect of quantile covariate related to lung and multiple metastases from the secondary inference was probably because the decline offset the growth within the τ interval $[0.1, 0.6]$ (Figure 3.2). Based on the test results, we might conclude that the survival difference between patients with lung and multiple metastatic breast cancer diseases was fluctuant over quantiles, and it reached to its minimum at 30^{th} quantile of HR-/HER2+ patients. Treatment strategies should be tailored based on the variation of prognostic effects.

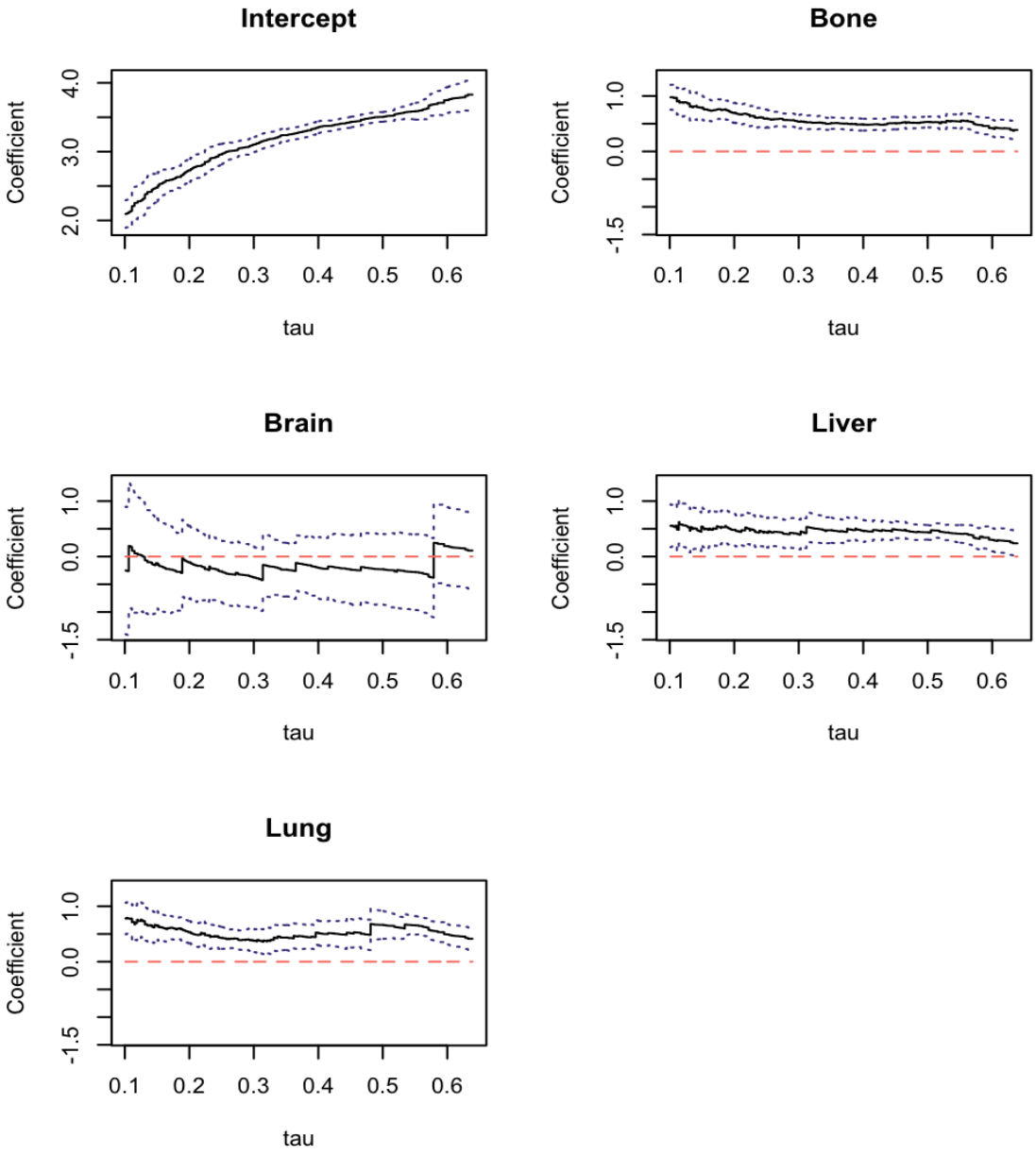


Figure 3. 2 Univariate analysis: estimated quantile coefficients with 95% pointwise confidence intervals in HR+/HER2- subtype for $\tau \in [0.1, 0.639]$

(2) HR+/HER2+ Subtype

Overall, HR+/HER2+ patients with metastasis in bone or liver organ had significantly better survival than those with metastasis in multiple organs over $\tau \in [0.1, 0.35]$ (P = 0.0002, P = 0.003; Table 3.6). As mentioned above in Cox PH regression, brain metastasis was excluded due to small sample size (n=7).

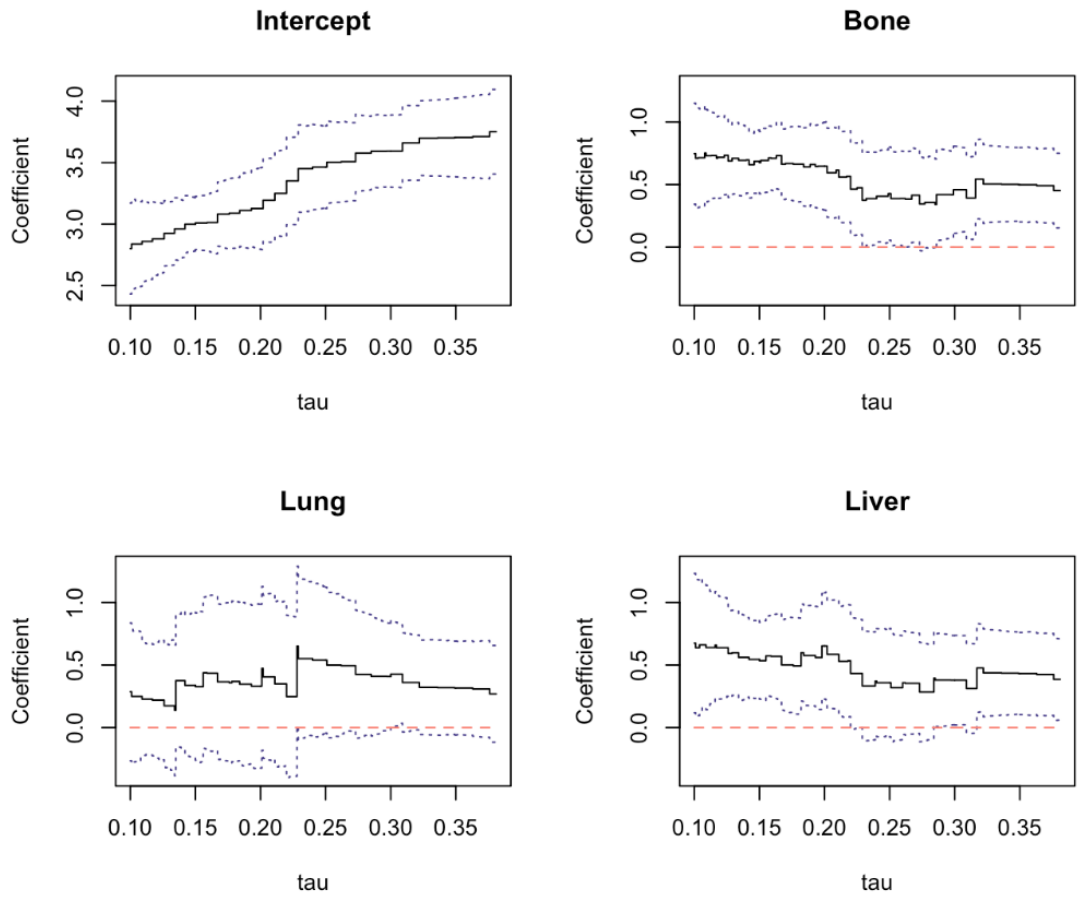


Figure 3. 3 Univariate analysis: estimated quantile coefficients with 95% pointwise confidence intervals in HR+/HER2+ subtype for $\tau \in [0.1, 0.381]$

In Figure 3.3, survival difference between bone and multiple metastases, or liver and metastasis had similar patterns over $\tau \in [0.1, 0.381]$: it appeared to be decreasing and gradually stable around 25^{th} quantile. Based on the Kaplan Meier curves for HR+/HER2- patients, fewer events happened at the end of the time and many patients lost to follow-up, particularly in patients with lung metastasis (Figure 3.3). Moreover, the sample size was small with relatively short follow-up time. Hence, a constancy test conducted within the τ interval $[0.1, 0.3]$ would be more powerful. For $\tau \in [0.1, 0.3]$, quantile effect of covariate related to bone and multiple metastasis was significantly inconsistent (P = 0.002) and the survival difference gradually vanished in this τ interval (Figure 3.3).

(3) HR-/HER2+ subtype

In HR-/HER2+ subtype, patients with bone, liver or lung metastasis had significantly better prognosis effect than those with multiple metastases when $\tau \in [0.1, 0.35]$ (P = 0.0497, P = 0.0002, P = 0.005; Table 3.6). However, these effects may be constant in $[0.1, 0.35]$ (Table 3.6).

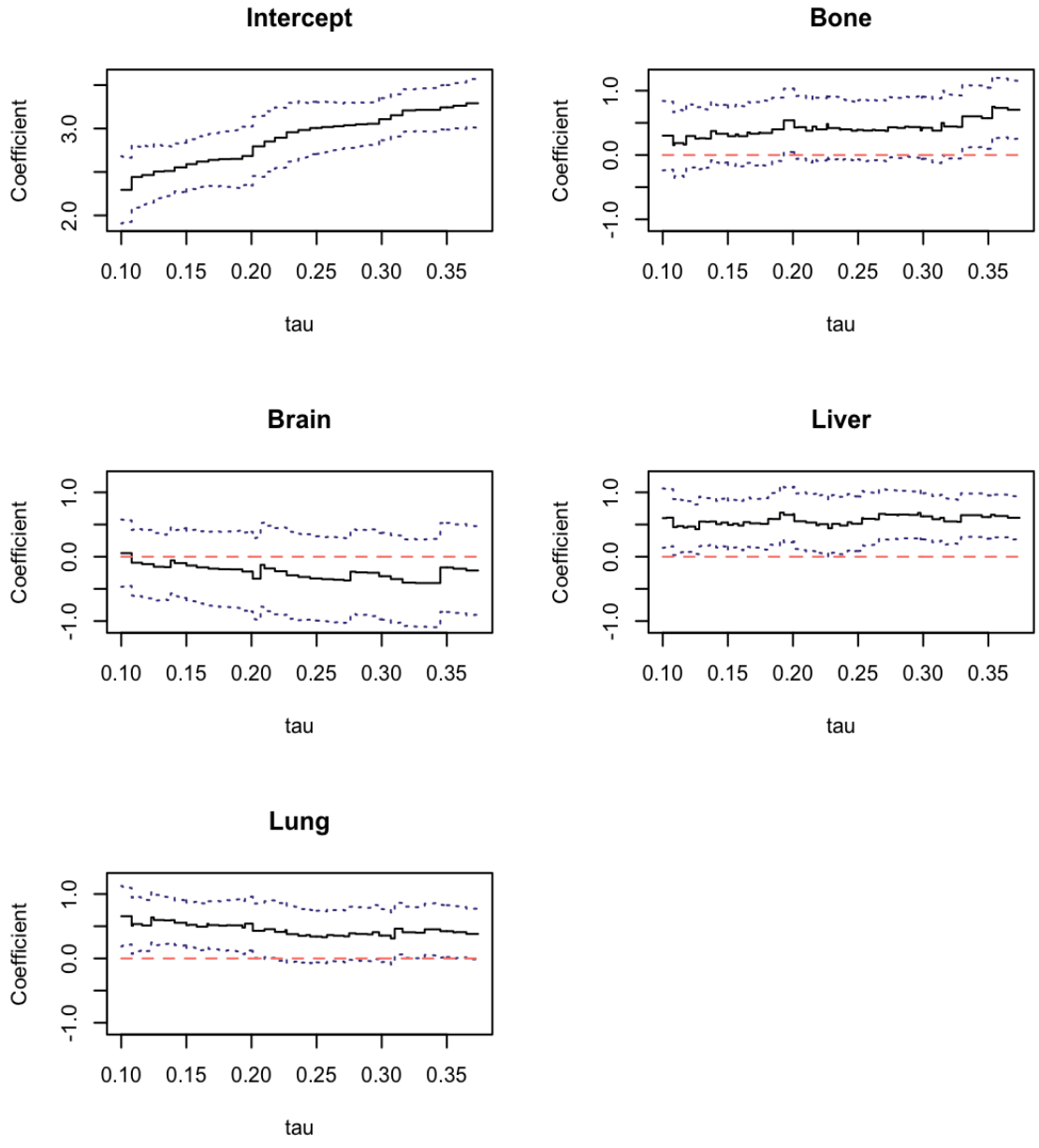


Figure 3. 4 Univariate analysis: estimated quantile coefficients with 95% pointwise confidence intervals in HR-/HER2+ subtype for $\tau \in [0.1, 0.374]$

(4) TNBC subtype

Overall tests showed significance in quantile effects related to bone, liver, lung versus multiple metastases within the interval $[0.1, 0.75]$. Patients with multiple metastatic breast cancer had significantly worse survival than those with single metastasis in bone, liver or lung organs (all P < .0001; Table 3.6). No overall effect on survival was shown between brain and multiple metastases (P = 0.414; Table 3.6). All covariate effects were constant over $\tau \in [0.1, 0.75]$. However, the positive quantile effect of bone metastasis versus multiple metastases, or lung metastasis versus multiple metastases may vary within the τ interval $[0.4, 0.75]$ (P < 0.0001, P = 0.074), possibly increasing in Figure 3.5.

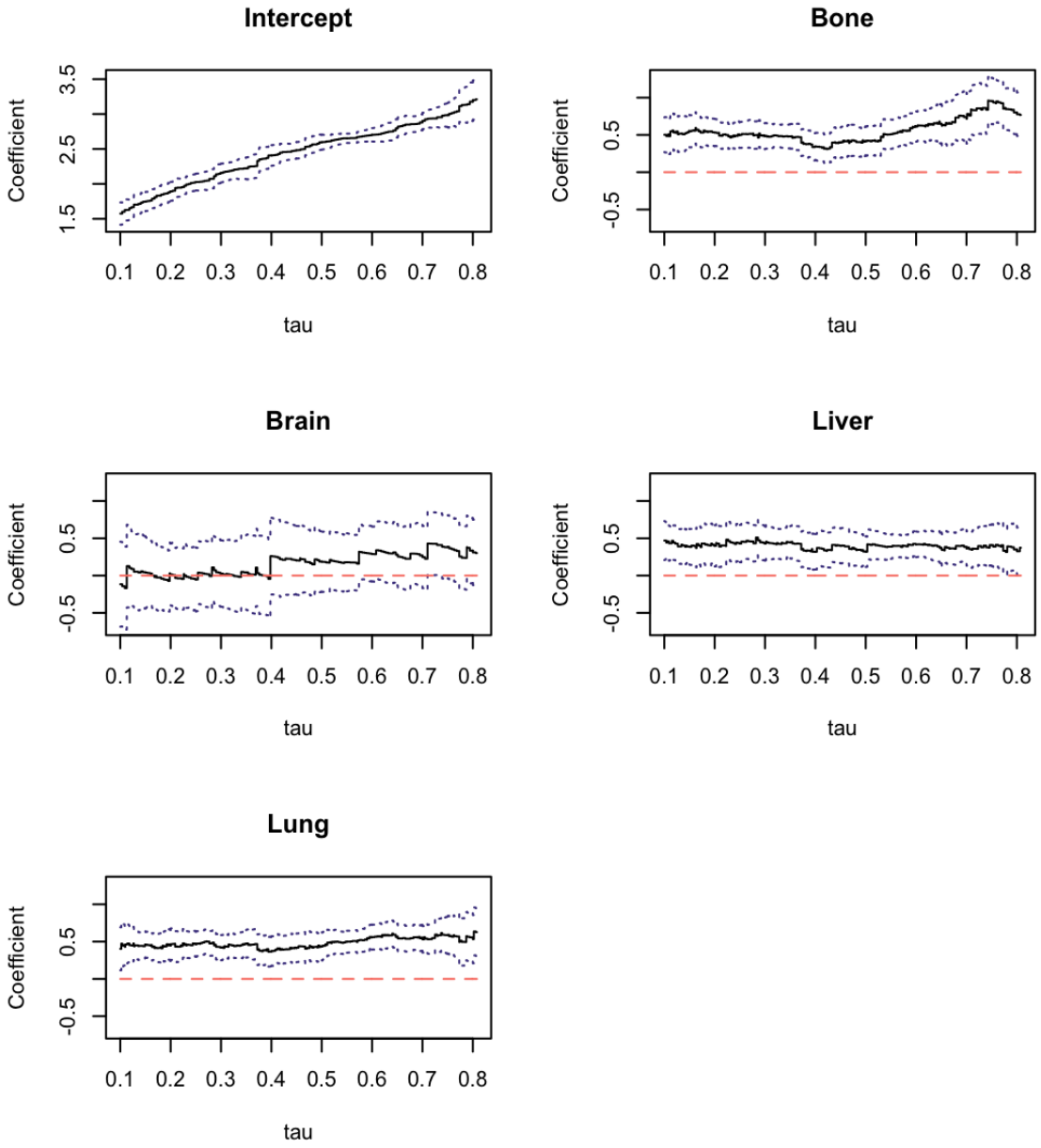


Figure 3. 5 Univariate analysis: estimated quantile coefficients with 95% pointwise confidence intervals in TNBC subtype for $\tau \in [0.1, 0.808]$

3.3.2 Multivariate Analysis

After adjusting for age, metastatic patterns associated with survival time were still different among breast cancer subtypes (Figure 3.6-3.9). Quantile covariate effects from multivariate regression models were either constant or varied. However, compared with univariate results, some of the patterns were changed in multivariate analyses. Parallel to univariate regression model, patients had metastases to multiple organs were considered as reference group. Table 3.7 summarized the average effects and test results of overall significance and constancy within specific quantile intervals from multivariate analyses. All subplots based on different reference categories of metastatic sites were shown in Appendix B.

Table 3. 7 Estimates of average covariate effects and results on hypothesis testing and second-stage inference in multivariate quantile regression model stratified by breast cancer subtype

	HR+/HER2-		
Metastatic Site	**Average Effect** $\tau\in[0.1,0.6]$	**Overall Significance** $\tau\in[0.1,0.6]$	**Constancy** $\tau\in[0.1,0.6]$
Bone	0.562	< 0.0001*	0.064
Brain	-0.210	0.339	0.752
Liver	0.373	0.0003*	0.414
Lung	0.585	< 0.0001*	0.586
Age	-0.017	< 0.0001*	0.004*
	HR+/HER2+		
Metastatic Site	**Average Effect** $\tau\in[0.1,0.4]$	**Overall Significance** $\tau\in[0.1,0.4]$	**Constancy** $\tau\in[0.1,0.4]$
Bone	0.545	0.0008*	0.121
Liver	0.441	0.034*	0.779
Lung	0.382	0.023*	0.241
Age	-0.021	0.0002*	0.322
	HR-/HER2+		
Metastatic Site	**Average Effect** $\tau\in[0.1,0.4]$	**Overall Significance** $\tau\in[0.1,0.4]$	**Constancy** $\tau\in[0.1,0.4]$
Bone	0.548	0.008*	0.214
Brain	-0.211	0.349	0.638
Liver	0.592	0.001*	0.453
Lung	0.532	0.003*	0.511
Age	-0.018	0.0005*	0.070
	TNBC		
Metastatic Site	**Average Effect** $\tau\in[0.1,0.8]$	**Overall Significance** $\tau\in[0.1,0.8]$	**Constancy** $\tau\in[0.1,0.8]$
Bone	0.572	< 0.0001*	0.056
Brain	0.141	0.426	0.182
Liver	0.400	< 0.0001*	0.586
Lung	0.495	< 0.0001*	0.124
Age	-0.007	0.002*	0.613

1. * Significant P value (<0.05).

2. Reference category: multiple metastases.

3. Average covariate effects were resampling based.

(1) HR+/HER2- subtype

Among patients with HR+/HER2- subtype, after adjusting for patient age, breast cancer metastasizing to multiple organs had more harmful effect on survival $\tau \in [0.1, 0.6]$, compared with breast cancer metastasizing to bone, liver or lung ($P < 0.0001$, $P = 0.0003$, $P < 0.0001$; Table 3.7). The survival difference between bone and multiple metastases had slightly decrease over quantiles, and showed marginally significance in variation ($P = 0.064$; Table 3.7, Figure 3.6).

Additionally, age had significantly negative effect on survival ($P < 0.0001$; Table 3.7) and it became less important on survival among patients who survive longer at larger quantiles ($P = 0.004$; Table 3.7). Except for age and bone metastasis, quantile effects for other covariates were all constant over $\tau \in [0.1, 0.6]$ (Table 3.7).

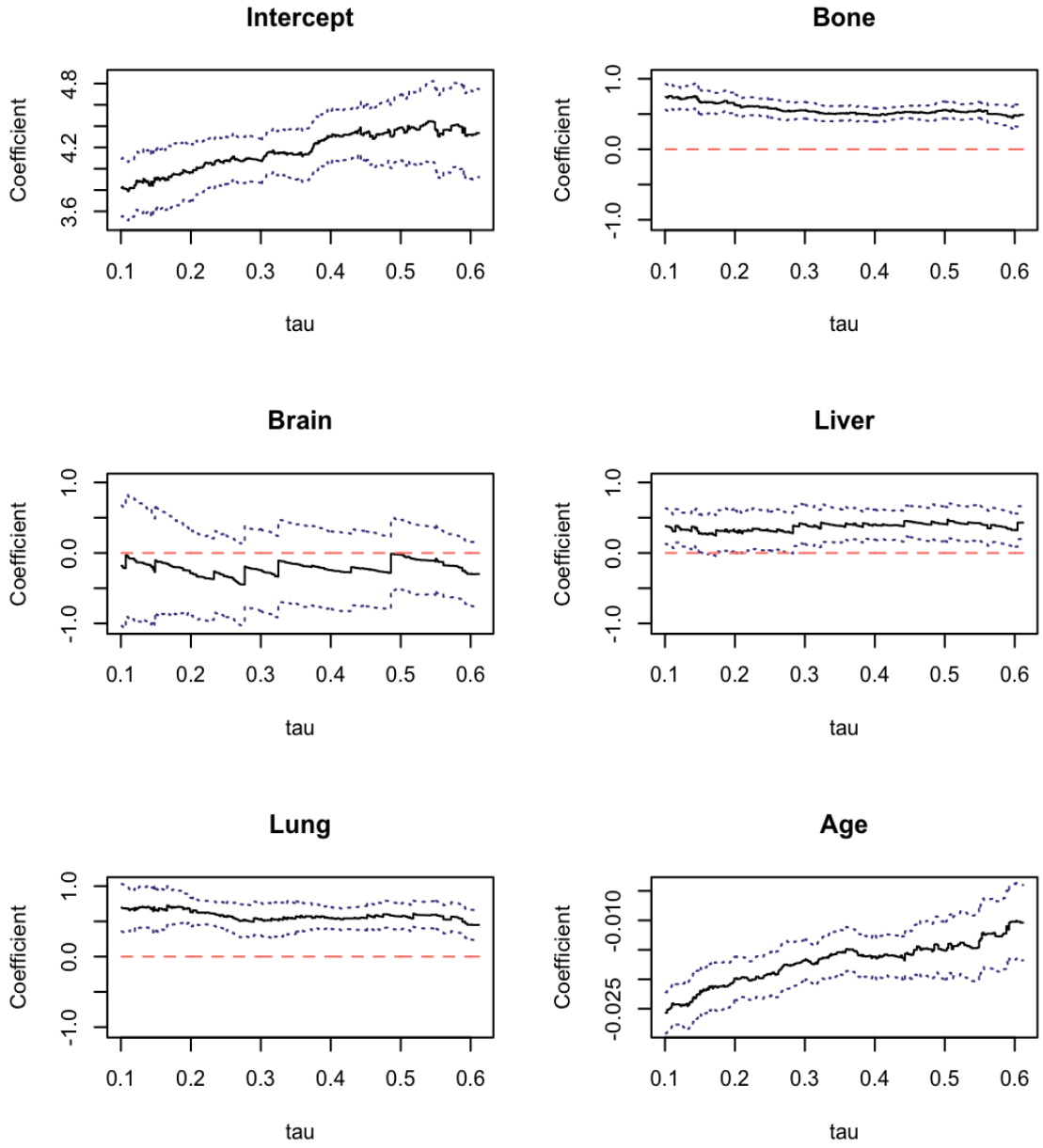


Figure 3. 6 Multivariate analysis: estimated quantile coefficients with 95% pointwise confidence intervals in HR+/HER2- subtype for $\tau \in [0.1, 0.612]$

(2) HR+/HER2+ subtype

Figure 3.7 displayed the quantile effects of metastatic sites and age in HR+/HER2+ breast cancer subtype. Patients with bone, lung, or liver metastasis had better overall prognosis than those with multiple metastases after controlling for age over $\tau \in [0.1, 0.4]$ (P = 0.0008, P = 0.034, P = 0.023,

P = 0.0002; Table 3.7). Age had negative effect on survival (P = 0.0002; Table 3.7). All the covariate effects were constant within this τ interval (Table 3.7).

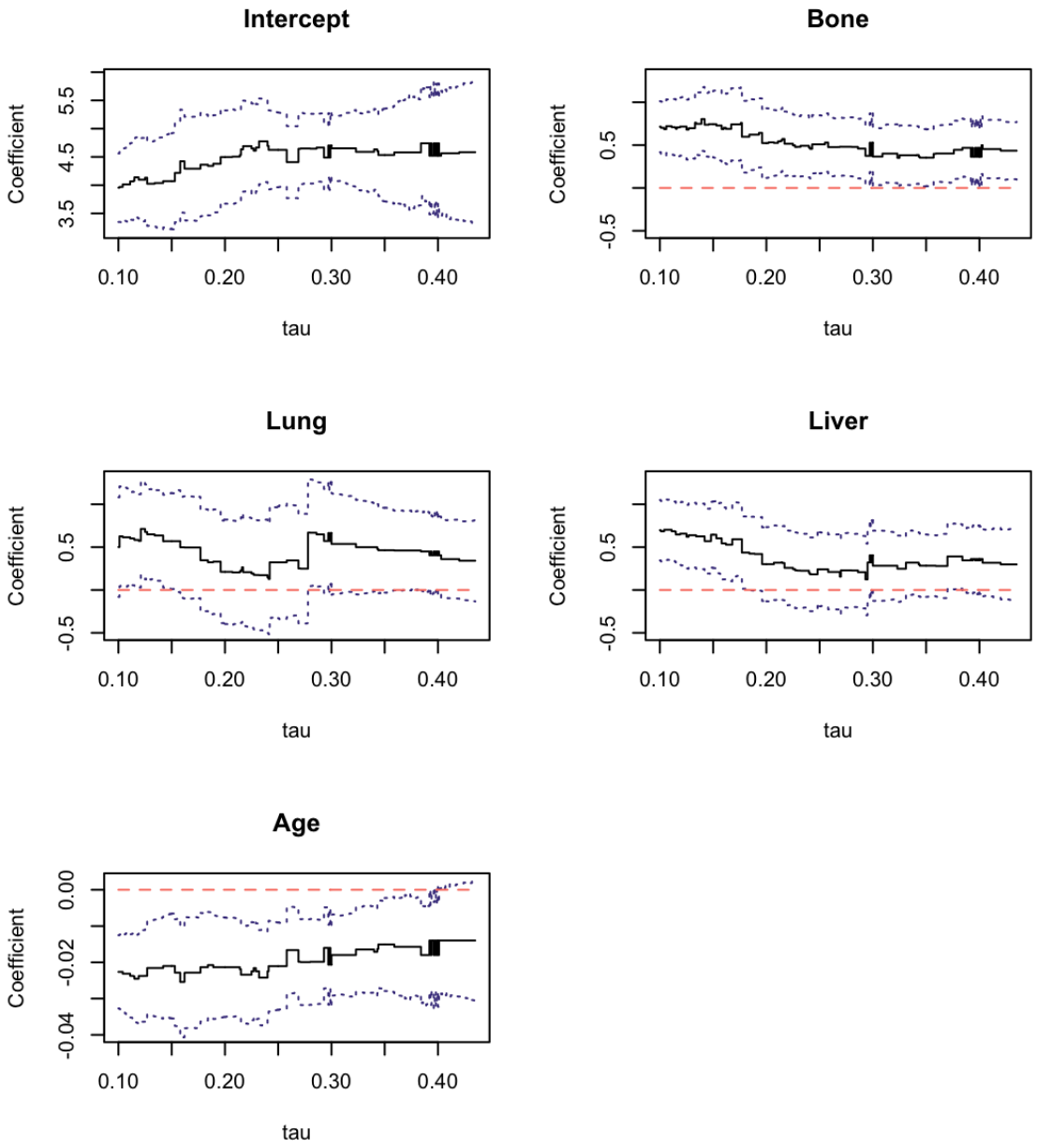


Figure 3. 7 Multivariate analysis: estimated quantile coefficients with 95% pointwise confidence intervals in HR+/HER2+ subtype for $\tau \in [0.1, 0.435]$

(3) HR-/HER2+ subtype

In HR-/HER2+ subtype, after adjusting for age, patients with bone, liver or lung oligometastasis had significantly better prognosis than patients with multiple metastases over $\tau \in [0.1, 0.4]$ (P = 0.008, P = 0.001, P = 0.003; Table 3.7).

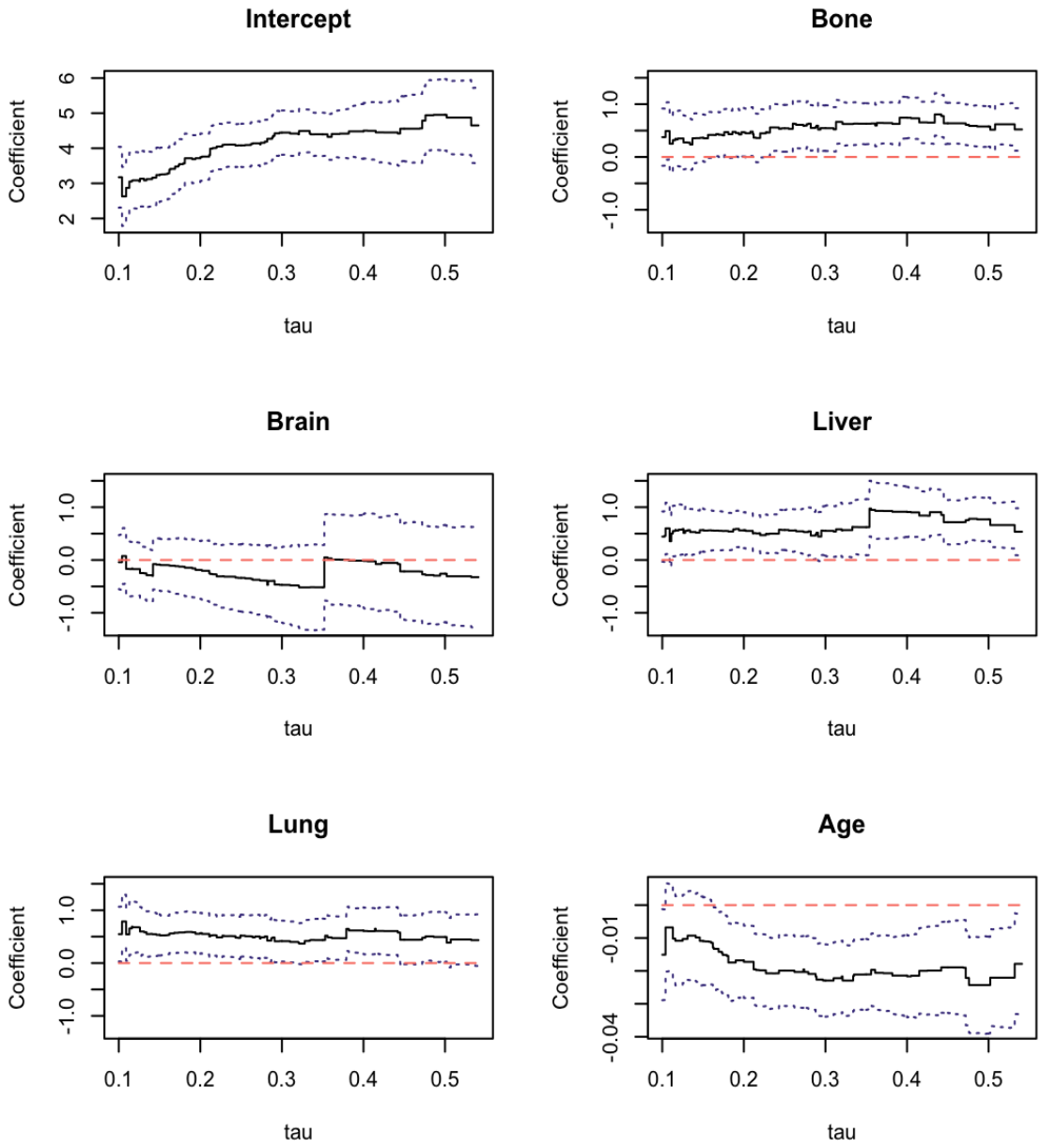


Figure 3. 8 Multivariate analysis: estimated quantile coefficients with 95% pointwise confidence intervals in HR-/HER2+ subtype for $\tau \in [0.1, 0.541]$

In Figure 3.8, age had negative quantile effect of prognosis (P = 0.0005; Table 3.7), which appeared to be decreasing as quantile increased (Figure 3.8). All these coefficient effects on survival were constant across quantiles (Table 3.7).

(4) TNBC subtype

After controlling for age, multiple metastases had significantly worse effect on survival than bone, liver or lung oligometastasis within quantile interval $[0.1, 0.8]$ (P < 0.0001, P = 0.426, P < 0.0001, P < 0.0001; Table 3.7). The estimated effect of age suggested that younger patients may have worse survival (P = 0.002; Table3.7, Figure 3.9) in the TNBC group.

The quantile effect of bone metastasis may vary over τ. The variation was only marginally significant (P = 0.056; Table 3.7). According to Figure 3.9, there was an increase in the coefficient effect of bone metastasis above 45^{th} quantile. A constancy test conducted within the τ interval [0.45, 0.8] showed significant survival variation between bone metastasis and multiple metastases over $\tau \in [0.45, 0.8]$ (P =0.0005). Within this quantile interval, the prognosis of TNBC patients with multiple metastases started to get worse and worse in comparison to those with metastasis to bone organ after receiving systemic treatments. Except for the bone metastasis, all the covariate effects were constant over $\tau \in [0.1, 0.8]$ (Table 3.7).

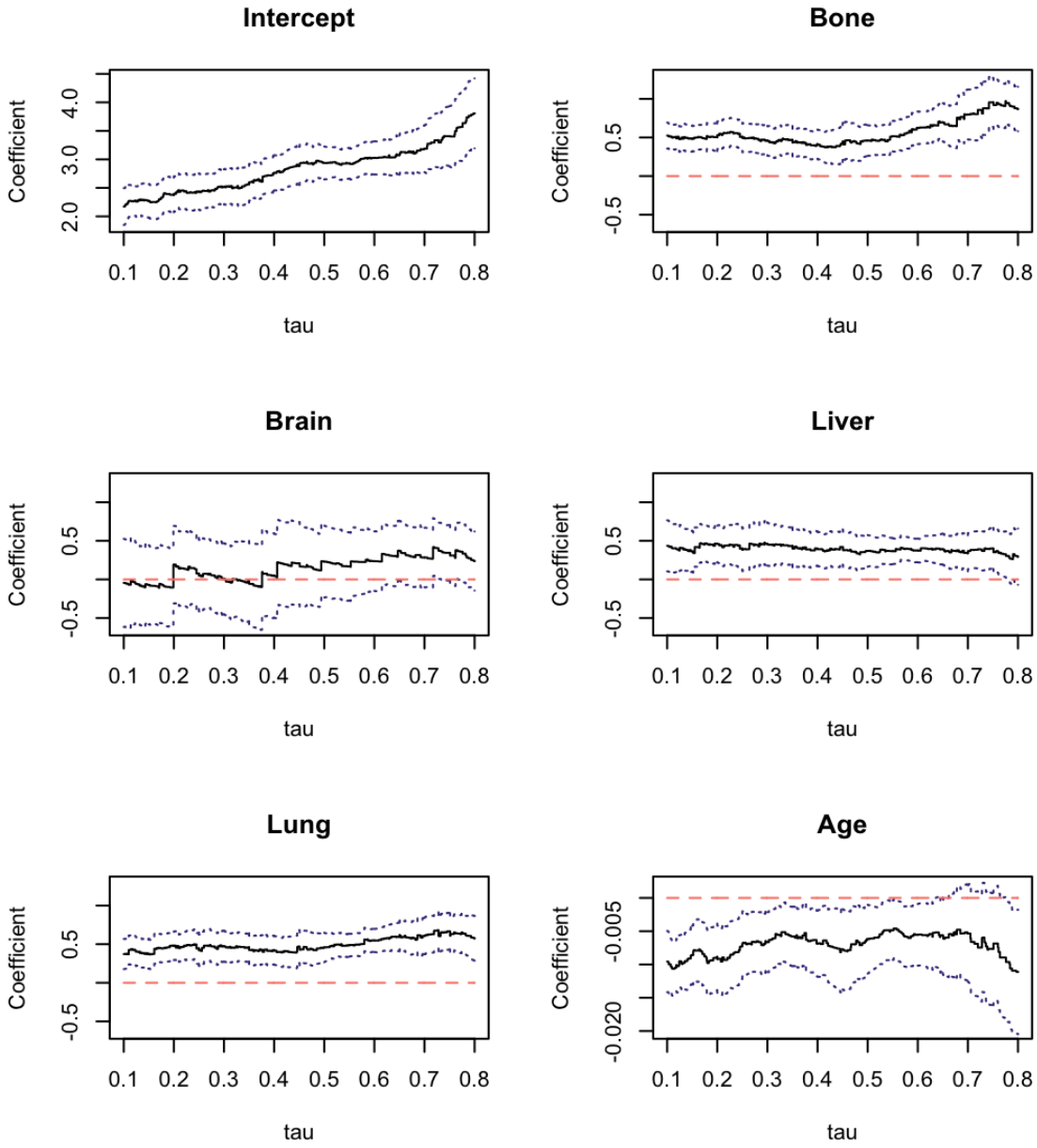


Figure 3. 9 Multivariate analysis: estimated quantile coefficients with 95% pointwise confidence intervals in TNBC subtype for $\tau \in [0.1, 0.827]$

Table 3.8 summarized the results from constancy tests among all significant overall quantile effects in Table 3.6 and Table 3.7. Patients with multiple metastases was the reference category for the covariates of metastatic sites in all quantile models. Based on the results from Table 3.6 and Table 3.7, patients who had metastasis in multiple organs never had significantly better prognosis than those who had single metastasis (bone, brain, liver, lung). Hence, patients with all the single metastases shown in Table 3.8 had significantly better overall survival than those with multiple metastases (reference group) in all breast cancer subtypes.

Table 3. 8 A summary of results on second stage inference based on quantile regression models

Significant Effect	Univariate Analysis		Multivariate Analysis	
	Constant	**Varying**	**Constant**	**Varying**
HR+/HER2-	Liver [0.1, 0.6]	Bone [0.1,0.6] Lung [0.1, 0.3] or [0.3, 0.55]	Bone, Liver, Lung [0.1, 0.6]	Age [0.1, 0.6]
HR+/HER2+	Liver [0.1, 0.3]	Bone [0.1, 0.3]	Bone, Liver, Lung, Age [0.1, 0.4]	--
HR-/HER2+	Bone, Liver, Lung [0.1, 0.35]	--	Bone, Liver, Lung, Age [0.1, 0.4]	--
TNBC	Bone, Liver, Lung [0.1, 0.75]	Bone [0.4, 0.75]	Bone, Liver, Lung, Age [0.1, 0.8]	Bone [0.45, 0.8]

1. Reference category: multiple metastases.

2. All single metastases (bone, brain, liver, lung) had better prognosis than multiple metastases.

CHAPTER IV: DISCUSSIONS

The purpose of this study was to determine the associations between metastatic site and overall survival in breast cancer patients with different breast cancer subtypes. A total of 5211 patients with metastatic breast cancer were selected from NCDB. The significant association between breast cancer metastatic site and subtype was examined. Two semiparametric regression models, including Cox proportional hazard model and censored quantile regression model, were used to evaluate the prognostic effects of metastatic sites in the same breast cancer subtype. Univariate and multivariate analyses were conducted under both models. Different association patterns between metastatic sites and overall survival were shown in different breast cancer subtypes.

In general, patients with bone metastasis appeared to have the best prognosis after receiving the systemic therapies, while patients with multiple metastases had the worst prognosis among all breast cancer subtypes, suggesting more aggressive therapies and medical care were needed. HER2+ patients or HER2- patients had similar metastatic patterns in prognosis regardless of variation over time. However, the results from these two semiparametric models were not completely the same. In Cox regression analyses, patients with single brain metastasis tended to have better overall survival than those with multiple metastases in TNBC subtype, while such a difference was not detected by censored quantile regressions. Notably, bone metastasis was more likely to have significant varying effects over time than other metastasis versus multiple metastases in univariate quantile regression analyses, except for HR-/HER2+ breast cancer. Among patients with HR+ breast cancer subtypes (HR+/HER2-, HR+/HER2+), the prognosis difference between single bone metastasis and multiple metastases became significantly smaller over time. While in TNBC patients, the prognosis difference started to become significantly larger around the median. In multivariate quantile regression analyses, age showed negative prognostic effect among patients with all subtypes and was varied in HR+/HER2- subtype. These varying quantile effects suggested more tailored treatments for patients were needed after receiving systemic adjuvant therapies.

4.1 Assumptions and Limitations

Even though different metastatic patterns were founded in different breast cancer subtypes, the study still carried assumptions and limitations, particularly from data source, sample selection and regression models.

4.1.1 Data Source and Patient Selection

National Cancer Database (NCDB) (2010-2013) with wider coverage was assumed to have much more representative results from the statistical analyses. However, this study had a relatively short follow-up time with a maximum of 71.23 months and a median of less than 30 months. It is mainly because of the update and expansion of NCDB in 2010 and short patient follow-up starting from 2015. For HR+ breast cancer patients who would have late metastatic events, the statistical results might be affected due to short follow-up time (Figure 3.1).

Additionally, NCDB only contained approximately 5% metastatic rate, which was much lower than the reported rate.[51] Furthermore, similar to other large database, NCDB had many incomplete records on sites of metastases. Although breast cancers metastasize to more than two organs with or without complete records were all considered as multiple metastases, some patients with only one metastasis record but missing information on other metastatic sites were all removed (n = 87). By removing these data, which did include patients with either single or multiple metastases, there is a chance that we introduced some bias into our analyses. Additionally, NCDB also had low rates in treatments, especially in HER2-targeted therapy among HER2+ patients. Among 1079 HER2+ patients, only 21.69 % of them received HER2-targeted therapy. Low overall rates of breast cancer metastasis and treatment and many missing data in NCDB reduced the sample size and likely resulted in some biases.

Another limitation of this study is the strict inclusion criteria that was used for patient selection. The impact of different therapies and sequence of therapies on survival among different breast

cancer subtypes is currently unknown. Moreover, clinical and molecular features, such as subtype and metastatic site, contribute to the heterogeneity in breast carcinomas.[52, 53] In an effort to avoid this heterogeneity, we made a criterion with the inclusion of only patients with surgery and systemic therapy. However, this strict inclusion criterion was also a limitation to some extent. While we were able to generate a more realistic picture of metastatic patterns for a specific group of patients with metastatic breast cancer, there were fewer covariates that needed to be adjusted for in multivariate analyses. Smaller differences would be expected between the results of univariate and multivariate analysis if the inclusion or exclusion criteria were too rigid.

4.1.2 Regression Models

(1) Cox proportional hazard model

In this study, Cox proportional models followed the assumptions of non-informative censoring and proportional hazard. However, the second assumption was easy to be violated in the real-world data. The survival difference between two types of metastatic breast cancers might vary at quantiles. In this study, the alternative approach (quantile regression analysis) had validated the inconstancy in part of the survival comparisons through second stage inferences. For instance, in univariate analysis, the positive effect of bone metastasis on survival became smaller compared with multiple metastases in HR-/HER2+ subtype over quantiles (P =.003; Table 3.6). After adjusting for age, the quantile effect still had small decrease (P = 0.064; Table 3.7).

In both univariate and multivariate analysis, the orders of overall survival affected by different sites of metastasis could not be confirmed through all pairwise comparisons, especially in TNBC subtype (Table 3.5). In other words, some of these hazard ratios that should have been different showed non-significance. Wide confidence intervals with little precision probably came from small sample sizes of subgroups or the violations of proportional hazard assumption from Cox model.

(2) Censored quantile regression model

Censored quantile regression model using Peng and Huang approach (2008) is under the assumption of conditionally independence censoring. Even though it reflects a more objective result over quantile rather than a constant effect from Cox model, it has its own limitations in this study.

First, sample size has large impact on determining whether the coefficient can be estimated at a specific quantile. In comparison to univariate model, more sample size is required to estimate $\boldsymbol{\beta}(\tau)$ over a relatively similar range of τ in multivariate analysis, especially when more categorical variables with multiple levels are included in the model. This issue with sample size becomes more severe when the follow-up time is shorter with more events might happened later.

Additionally, the accuracy of estimates varies over quantile levels**.** In general, there are few events happened at the extreme quantiles (τ close to 0 or 1). The estimated effects close to the extreme usually are less stable and accurate, compared with those close to the median (0.5).[54] This is mainly because of the distribution of time-to-event data and grid-based method for quantile estimation[43]. To gain more reliable results, we only considered the covariate effects for $\tau \in [0.1, \tau_u]$ in this study, where $\tau_u < \tau_U$, τ_U was the maximum τ obtained from a quantile model estimation.

In this study, there were only a small number of patients with single brain metastasis in all subtype (HR+/HER2-: 27, HR+/HER2+: 7, HR-/HER2+: 14, TNBC: 35; Table 3.1). Moreover, the patient follow-up time with a median of less than 30 months. According to these data limitations, we might only acquire quantile estimates in relatively short range with less precision and accuracy.

(3) Conclusions of the two models

Survival differences assessed by Cox regression model were under the strong assumption of proportionality, which would cause bias estimations sometimes. Even though censored quantile regression was a more effective approach in assessing the variations of effects on survival, it needed

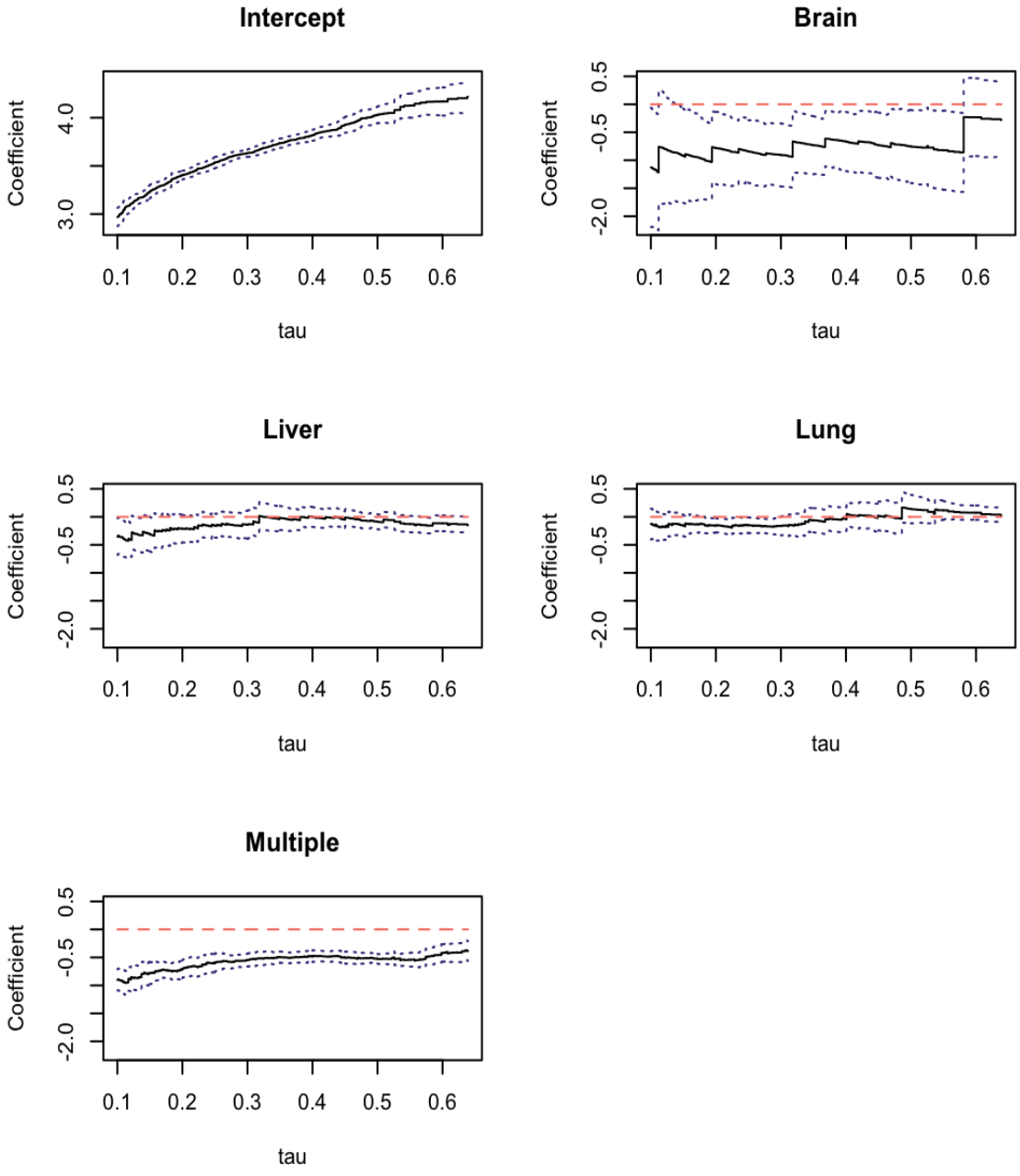
larger sample size to satisfy its accuracy, especially at the extreme quantiles (e.g. τ close to 0 or 1). In practice, one may conduct Cox PH regression as part of the primary analysis. When the PH assumption is dubious, or when the interest lies in covariate effects on survival times themselves, censored quantile regression may be performed as an alternative analysis which can be easily and stably implemented and provide straightforward physical interpretations.

4.2 Future Research

As mentioned previously, breast cancer is highly heterogeneous.[18,19] To reflect more realistic patterns, further work may focus on a specific breast cancer subtype, such as TNBC. TNBC patients lack the gene expressions of ER, PR, and HER2 receptors.[12-14] Accordingly, targeted therapies, such as immunotherapy, and HER2-targeted therapy, were usually ineffective among these patients. Patients with TNBC usually have worse survival than other breast cancer subtypes after distant metastasis. Based on the available therapy information from NCDB, it would be interesting to conduct a more detailed investigation on how chemotherapy or radiation affect the survival among TNBC patients with metastatic breast cancer using quantile regression models.

Appendix A: Regression Quantiles in Univariate Analysis

A.1 HR+/HER2- Subtype

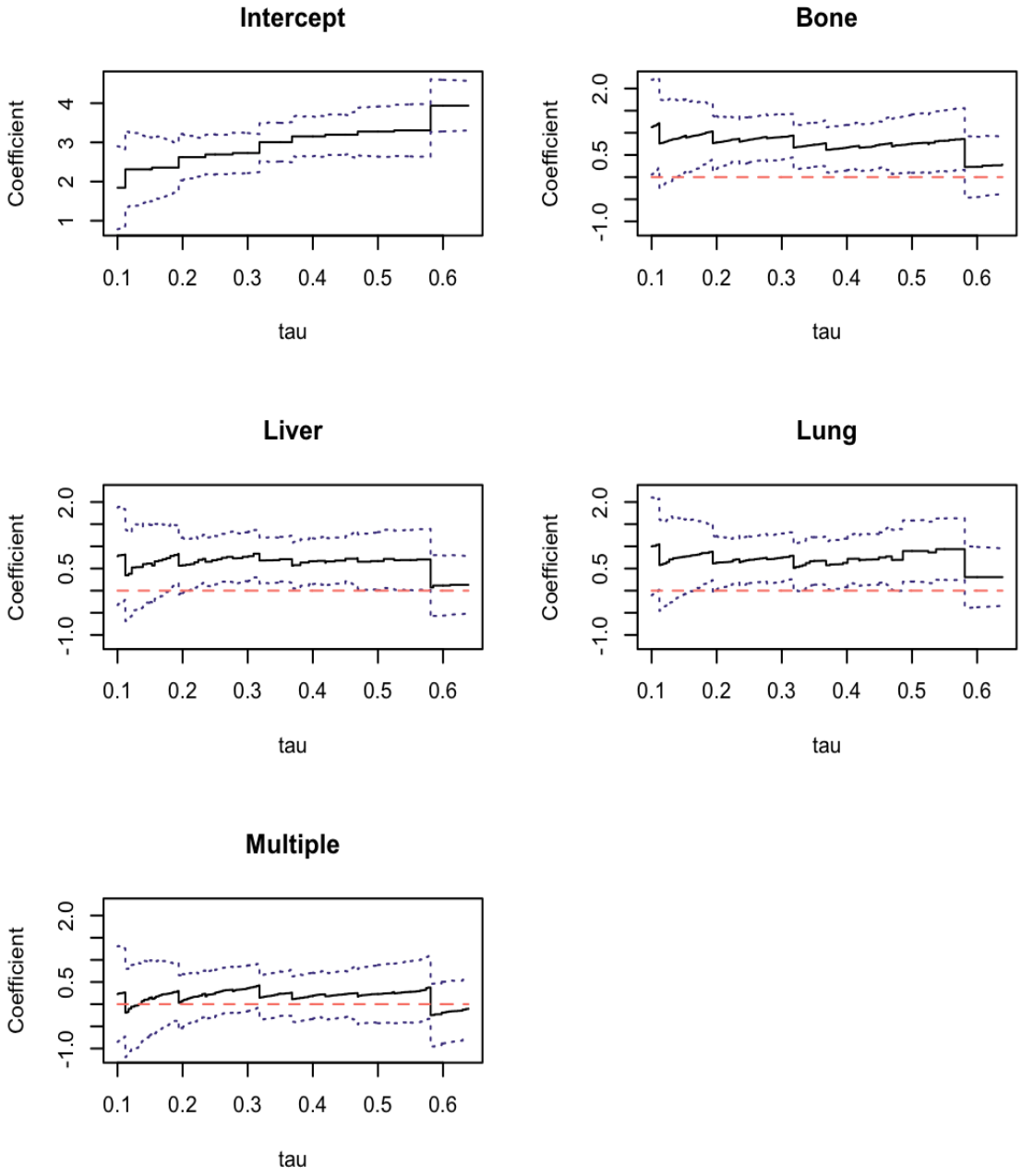


(A) Bone Metastasis

1. $\tau \in [0.1, 0.639]$.

2. Reference category of metastasis covariate: bone.

A.1 HR+/HER2- Subtype

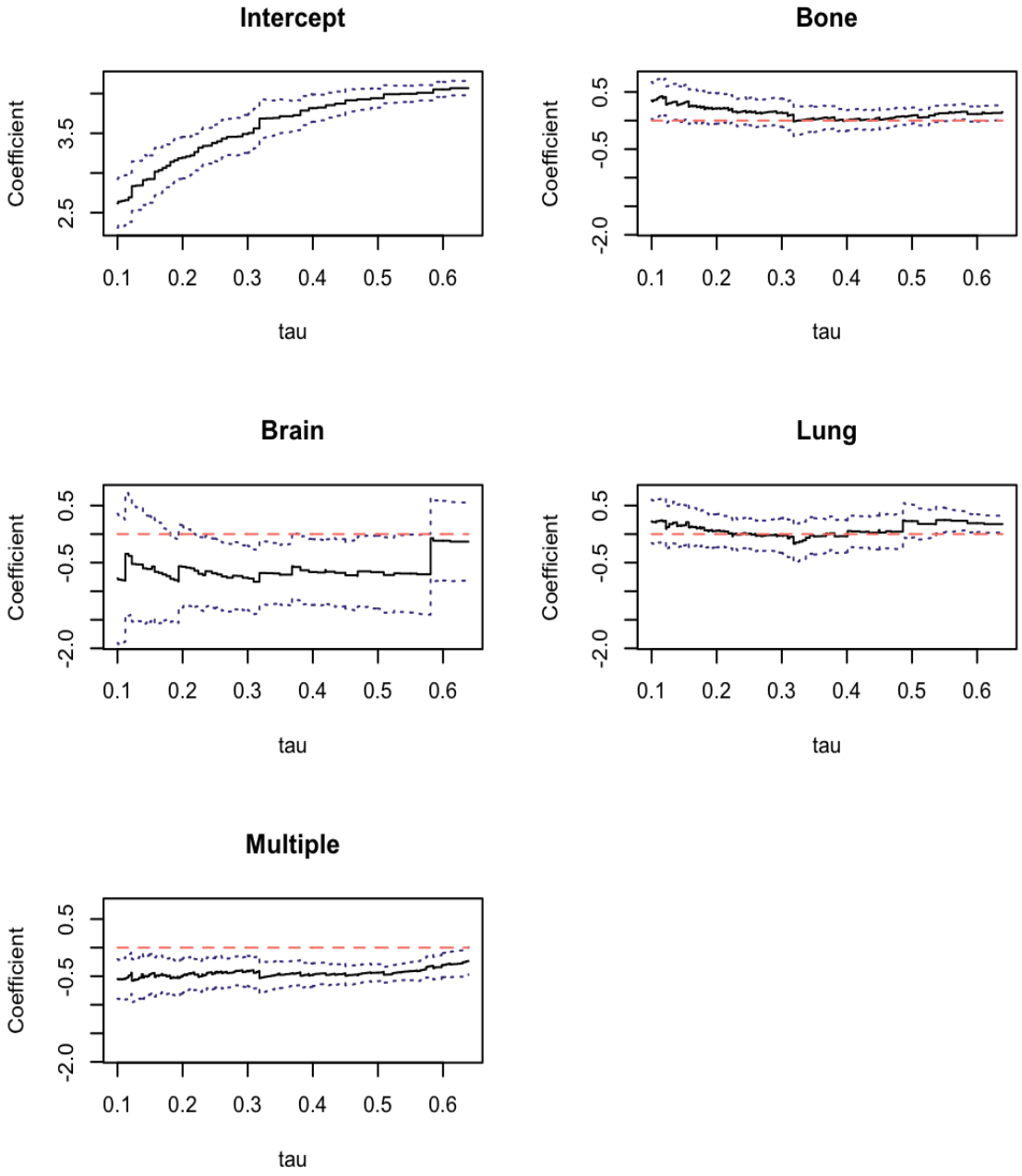


(B) Brain Metastasis

1. $\tau \in [0.1, 0.639]$.

2. Reference category of metastasis covariate: brain.

A.1 HR+/HER2- Subtype

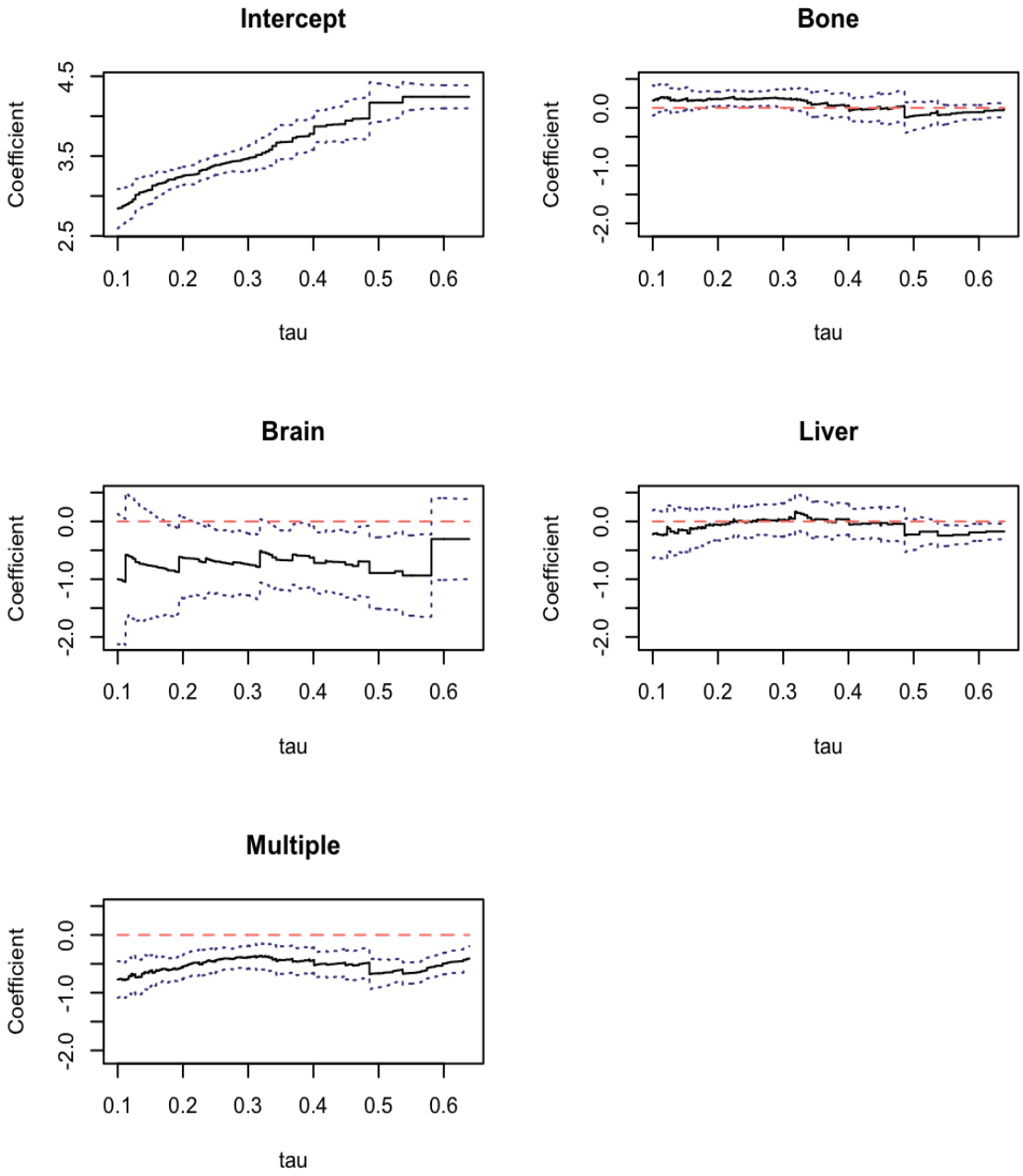


(C) Liver Metastasis

1. $\tau \in [0.1, 0.639]$.

2. Reference category of metastasis covariate: liver.

A.1 HR+/HER2- Subtype

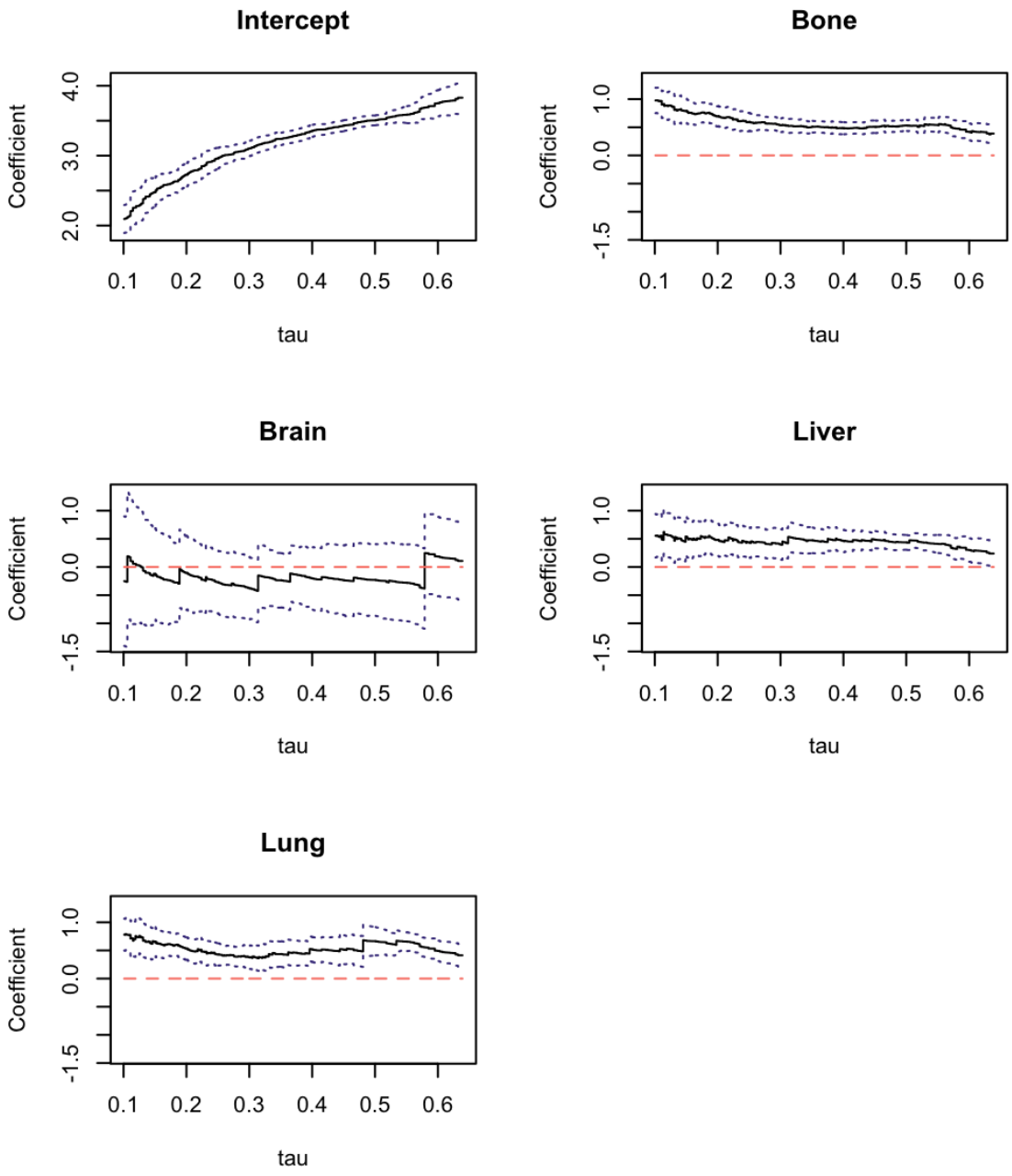


(D) Lung Metastasis

1. $\tau \in [0.1, 0.639]$.

2. Reference category of metastasis covariate: lung.

A.1 HR+/HER2- Subtype

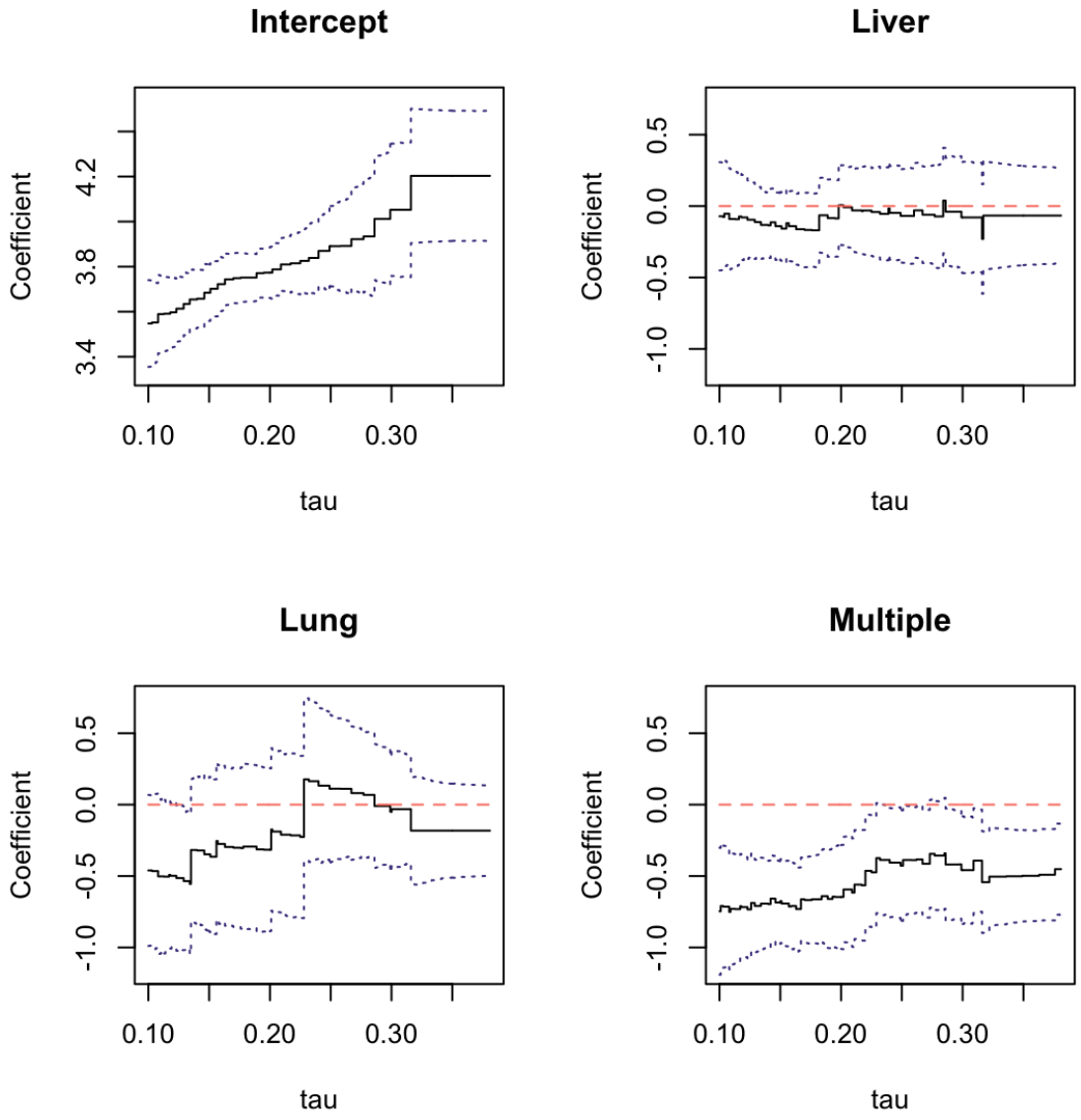


(E) Multiple Metastases

1. $\tau \in [0.1, 0.639]$.

2. Reference category of metastasis covariate: multiple.

A.2 HR+/HER2+ Subtype

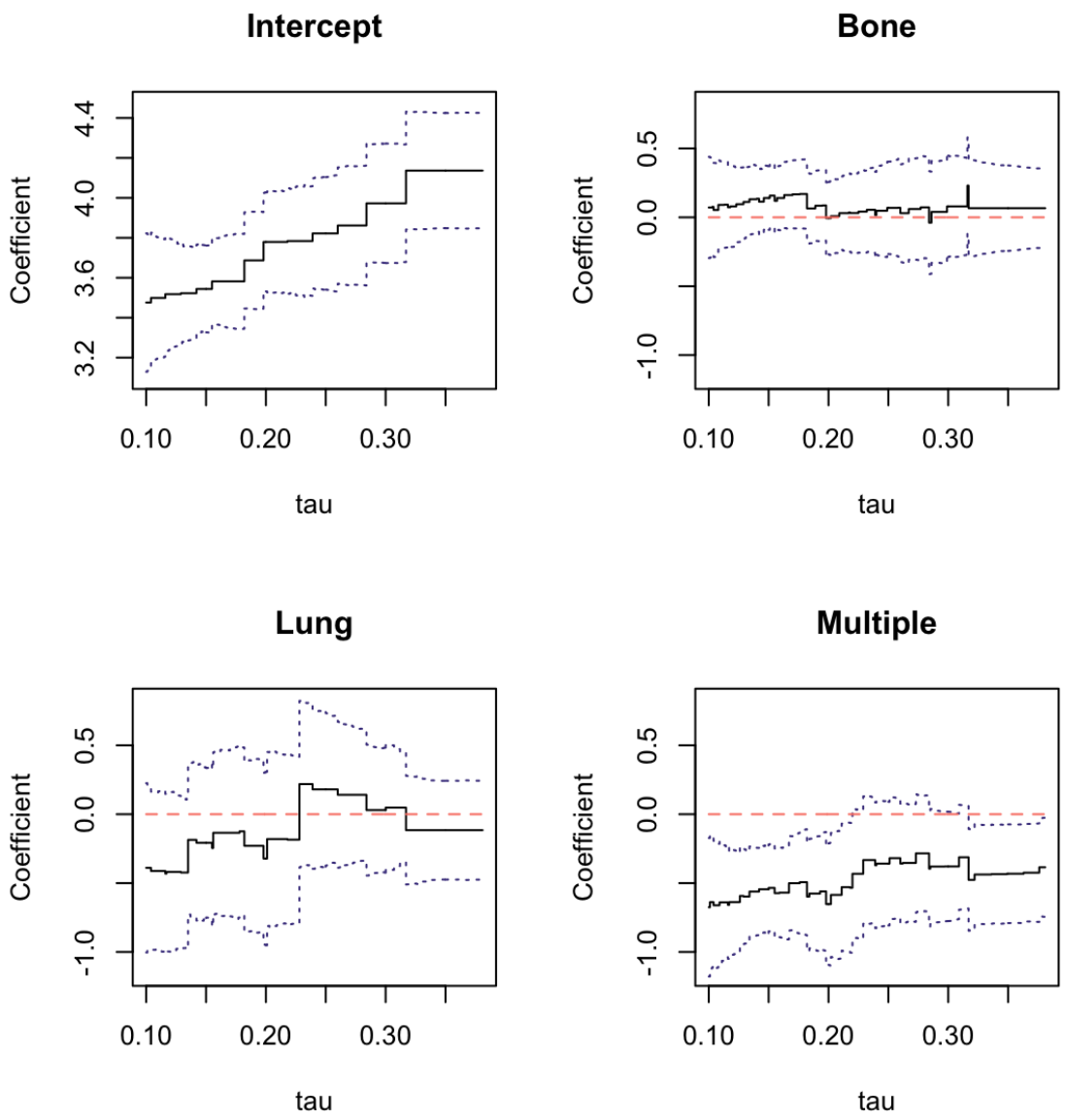


(A) Bone Metastasis

1. $\tau \in [0.1, 0.381]$.

2. Reference category of metastasis covariate: bone.

A.2 HR+/HER2+ Subtype

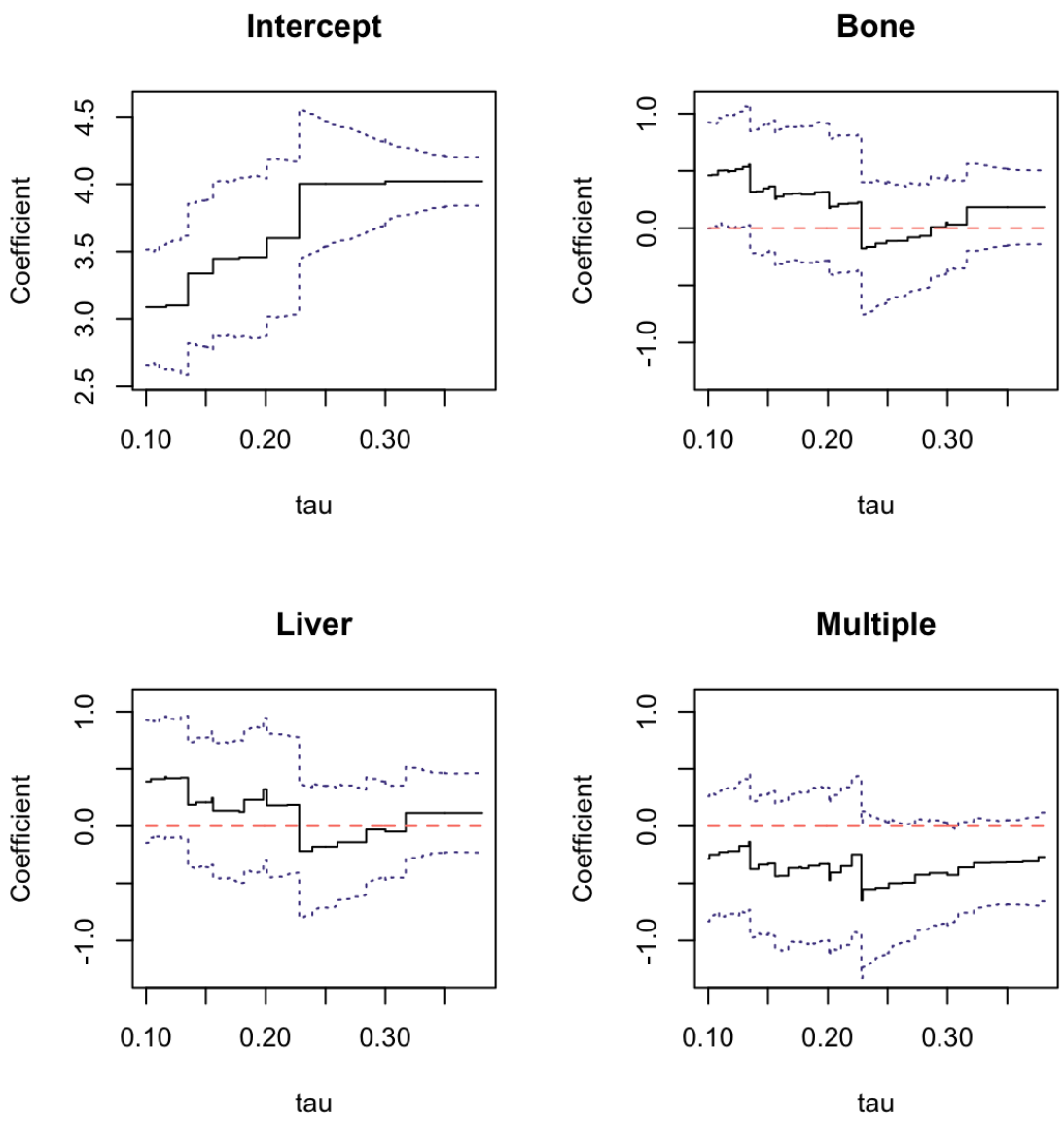


(B) Liver Metastasis

1. $\tau \in [0.1, 0.381]$.

2. Reference category of metastasis covariate: liver.

A.2 HR+/HER2+ Subtype

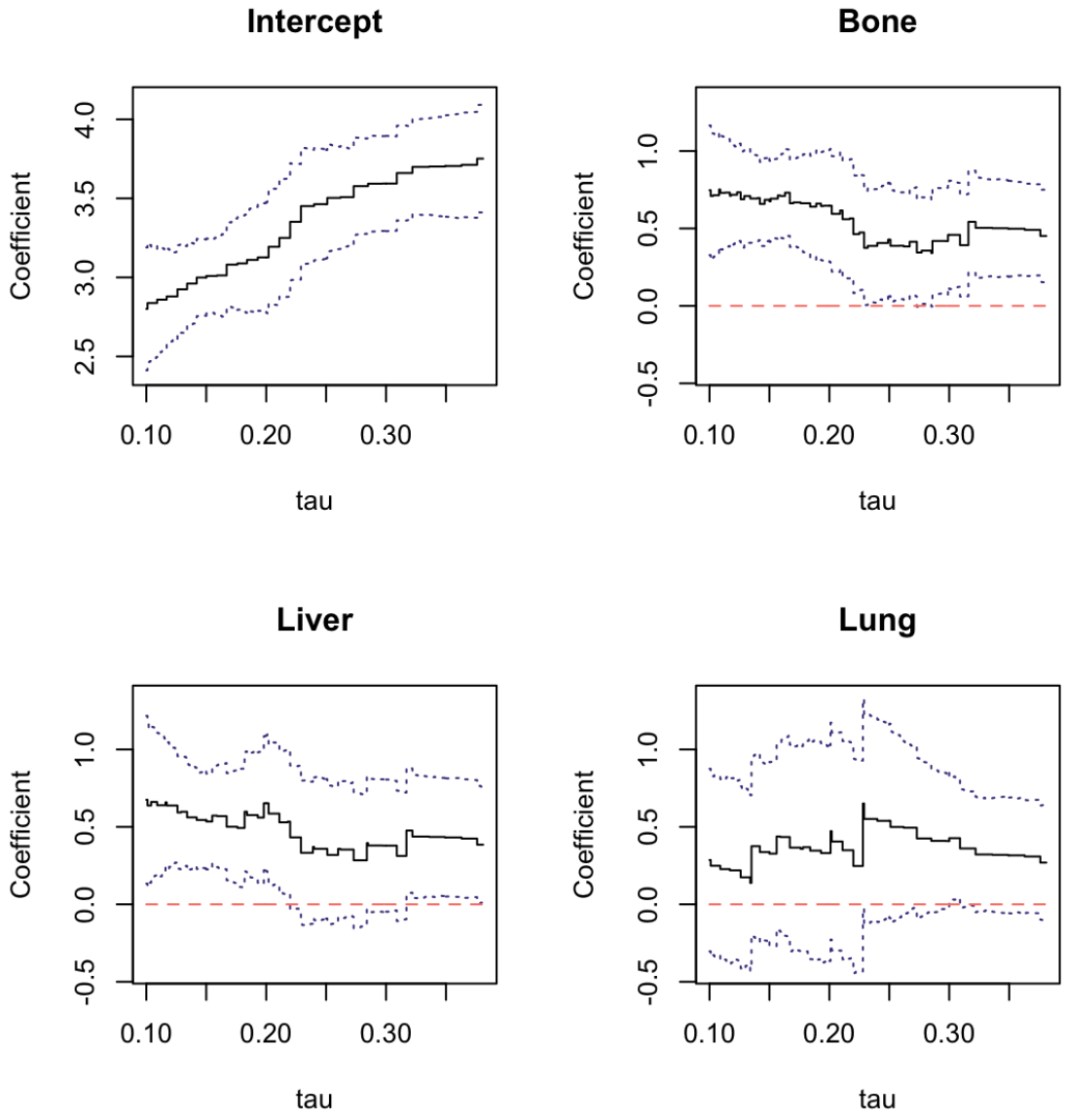


(C) Lung Metastasis

1. $\tau \in [0.1, 0.381]$.

2. Reference category of metastasis covariate: lung.

A.2 HR+/HER2+ Subtype

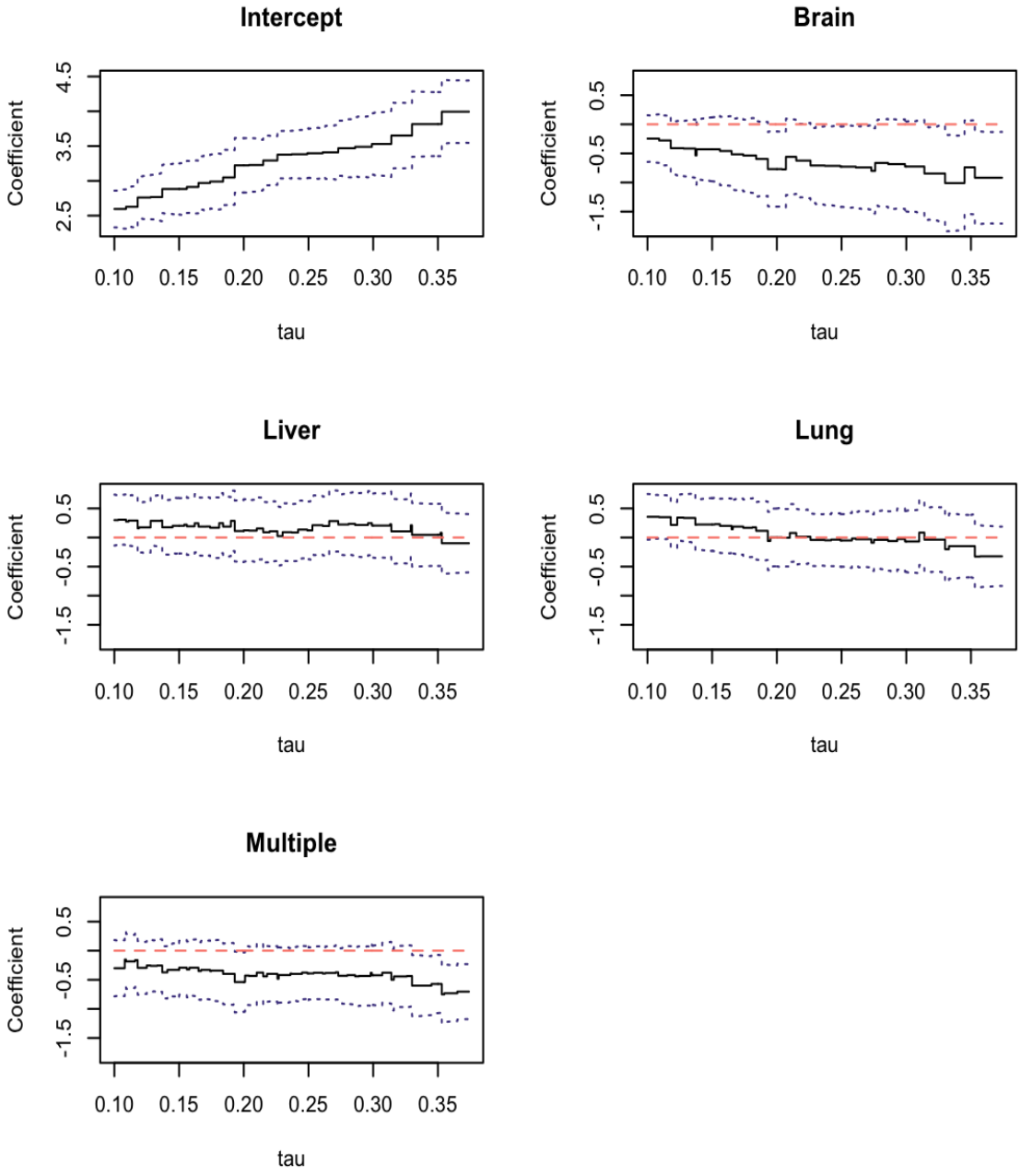


(D) Multiple Metastasis

1. $\tau \in [0.1, 0.381]$.

2. Reference category of metastasis covariate: multiple.

A.3 HR-/HER2+ Subtype

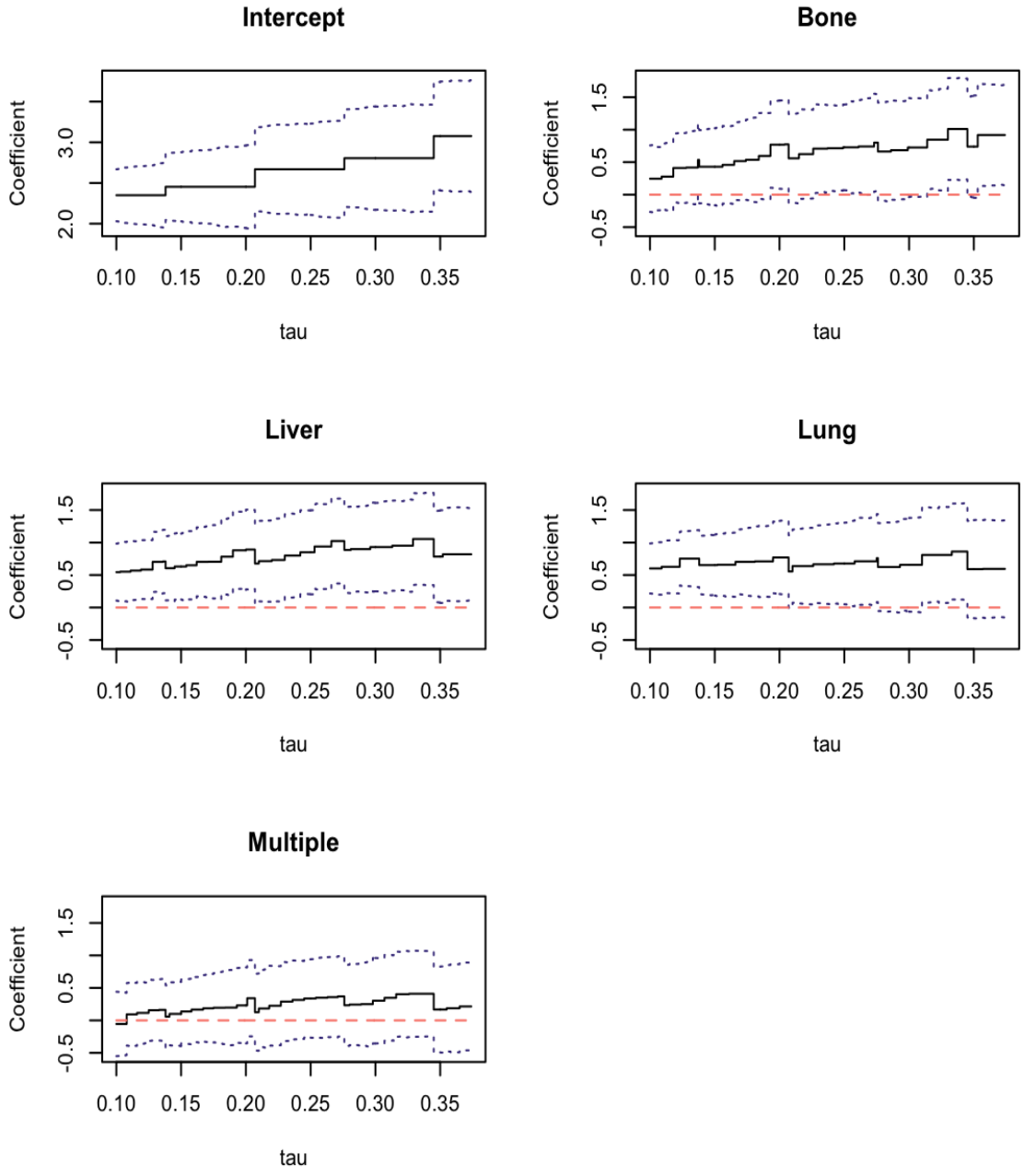


(A) Bone Metastasis

1. $\tau \in [0.1, 0.374]$.

2. Reference category of metastasis covariate: bone.

A.3 HR-/HER2+ Subtype

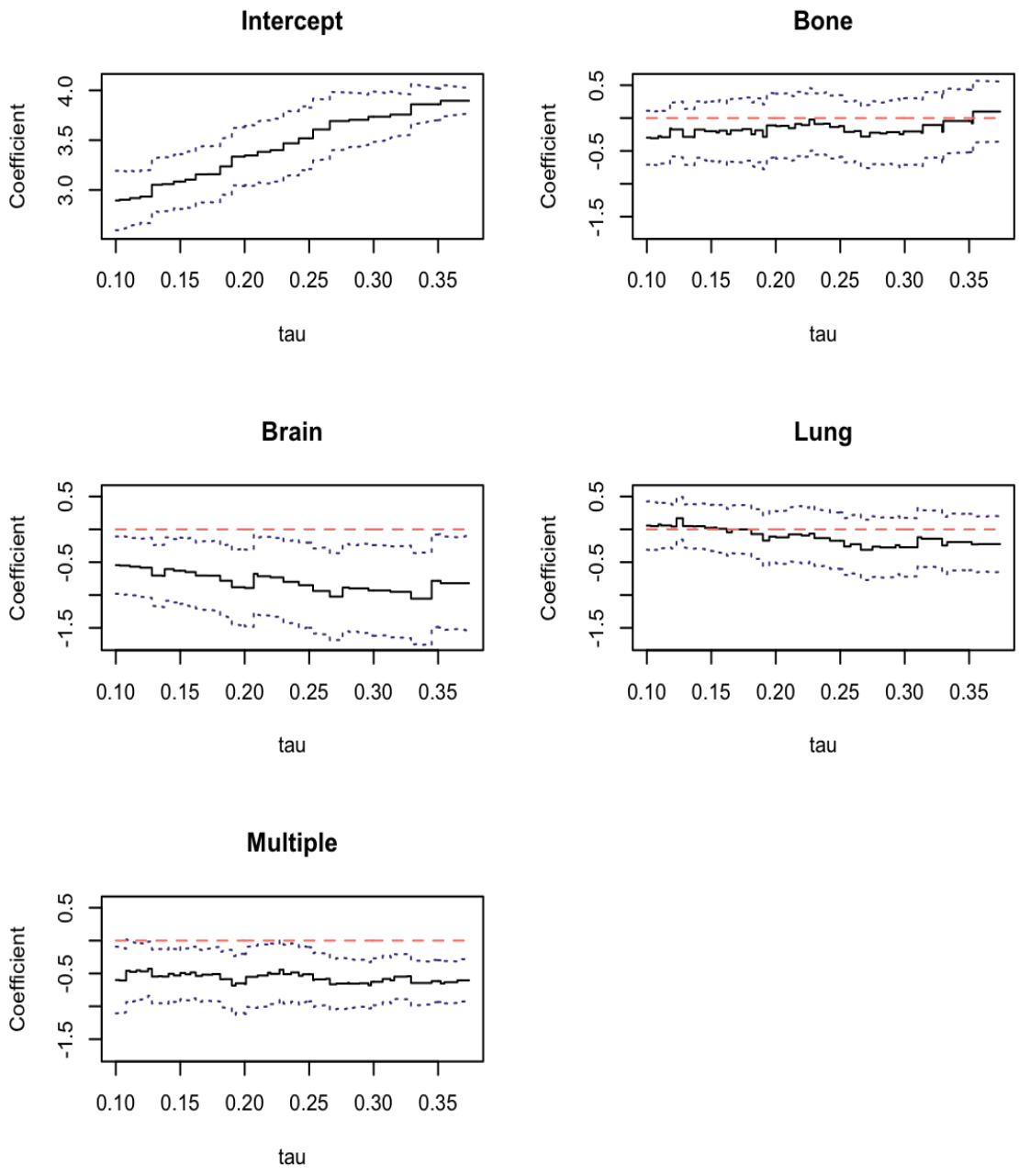


(B) Brain Metastasis

1. $\tau \in [0.1, 0.374]$.

2. Reference category of metastasis covariate: brain.

A.3 HR-/HER2+ Subtype

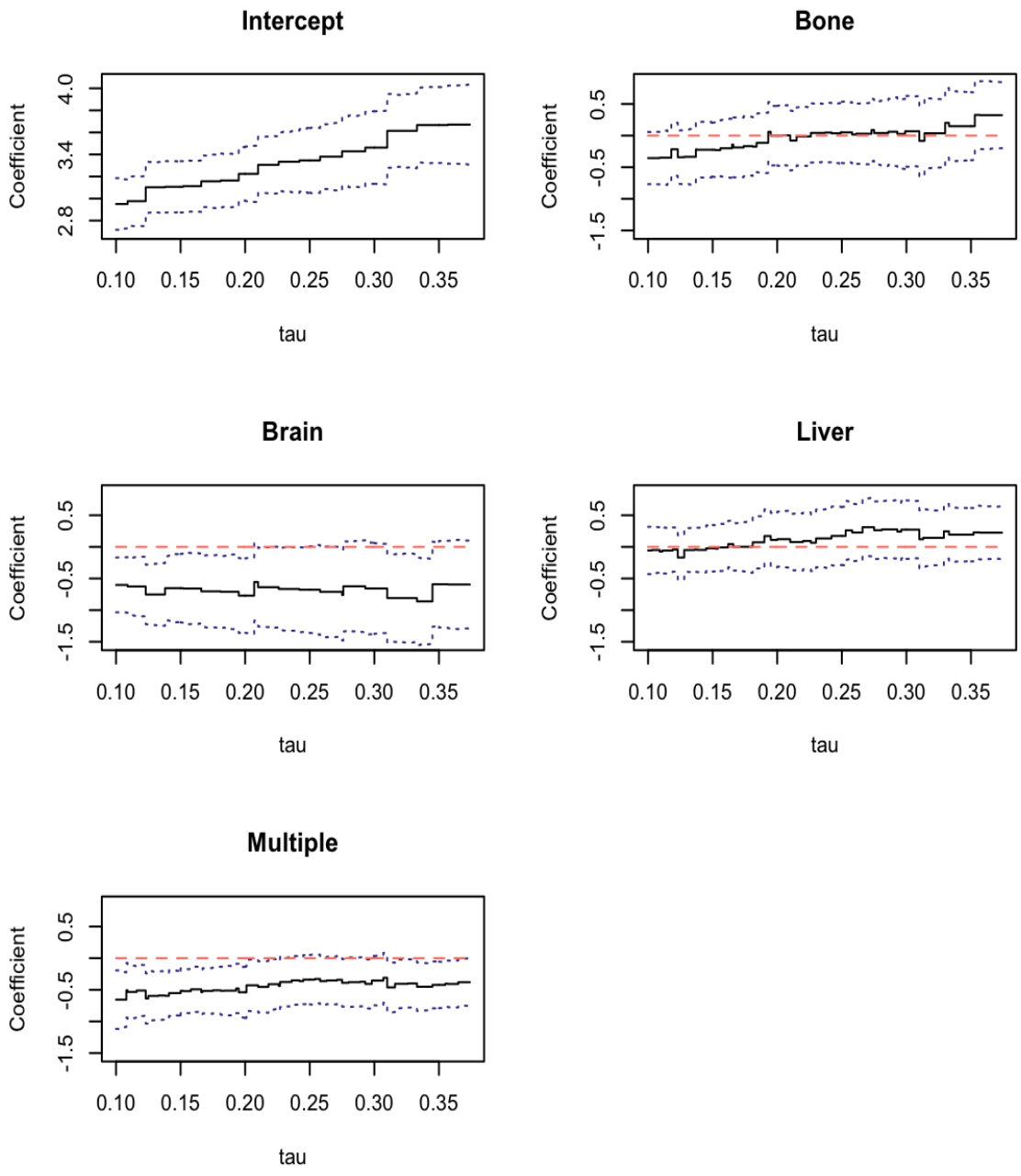


(C) Liver Metastasis

1. $\tau \in [0.1, 0.374]$.

2. Reference category of metastasis covariate: liver.

A.3 HR-/HER2+ Subtype

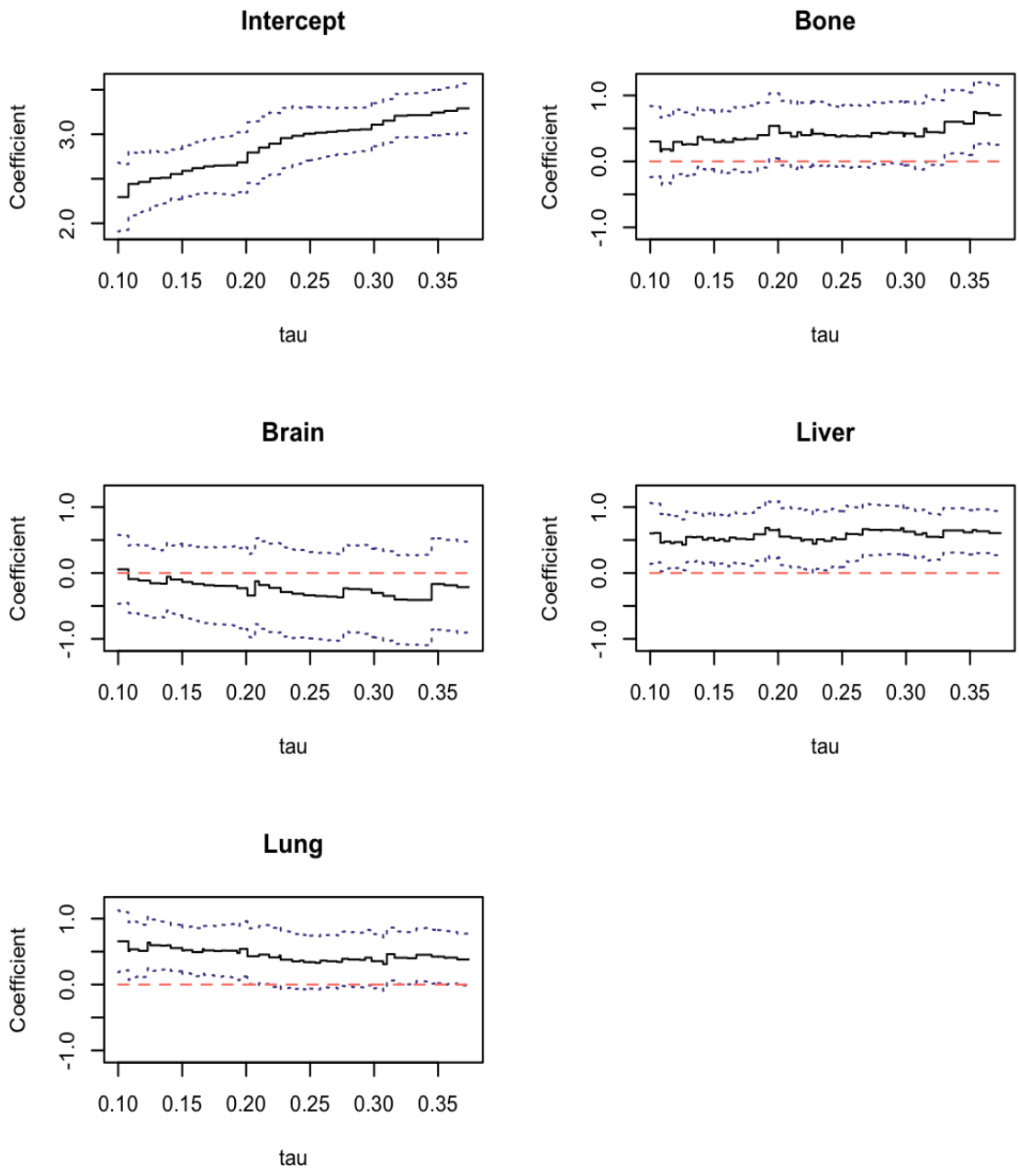


(D) Lung Metastasis

1. $\tau \in [0.1, 0.374]$.

2. Reference category of metastasis covariate: lung.

A.3 HR-/HER2+ Subtype

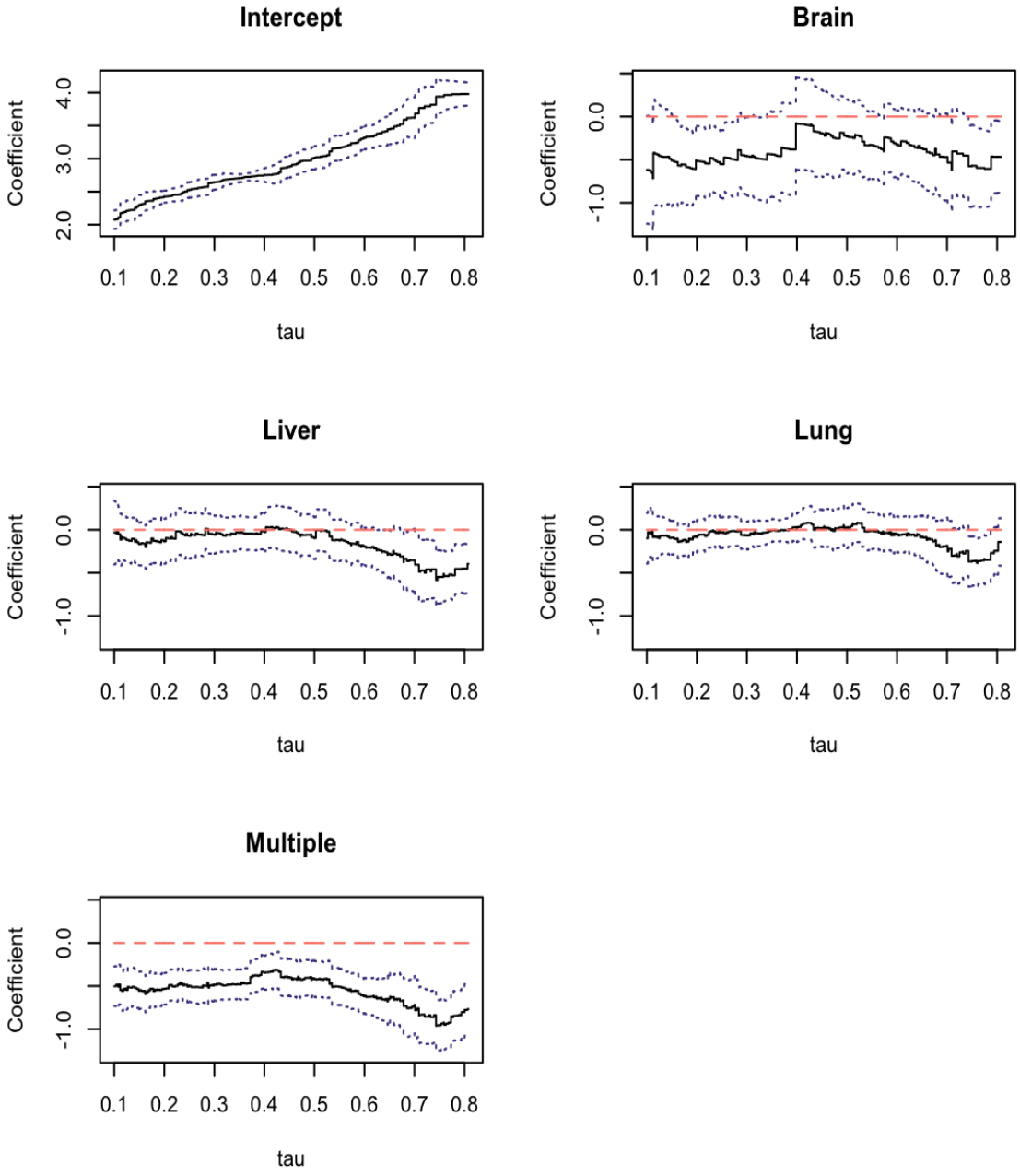


(E) Multiple Metastases

1. $\tau \in [0.1, 0.374]$.

2. Reference category of metastasis covariate: multiple.

A.4 TNBC Subtype

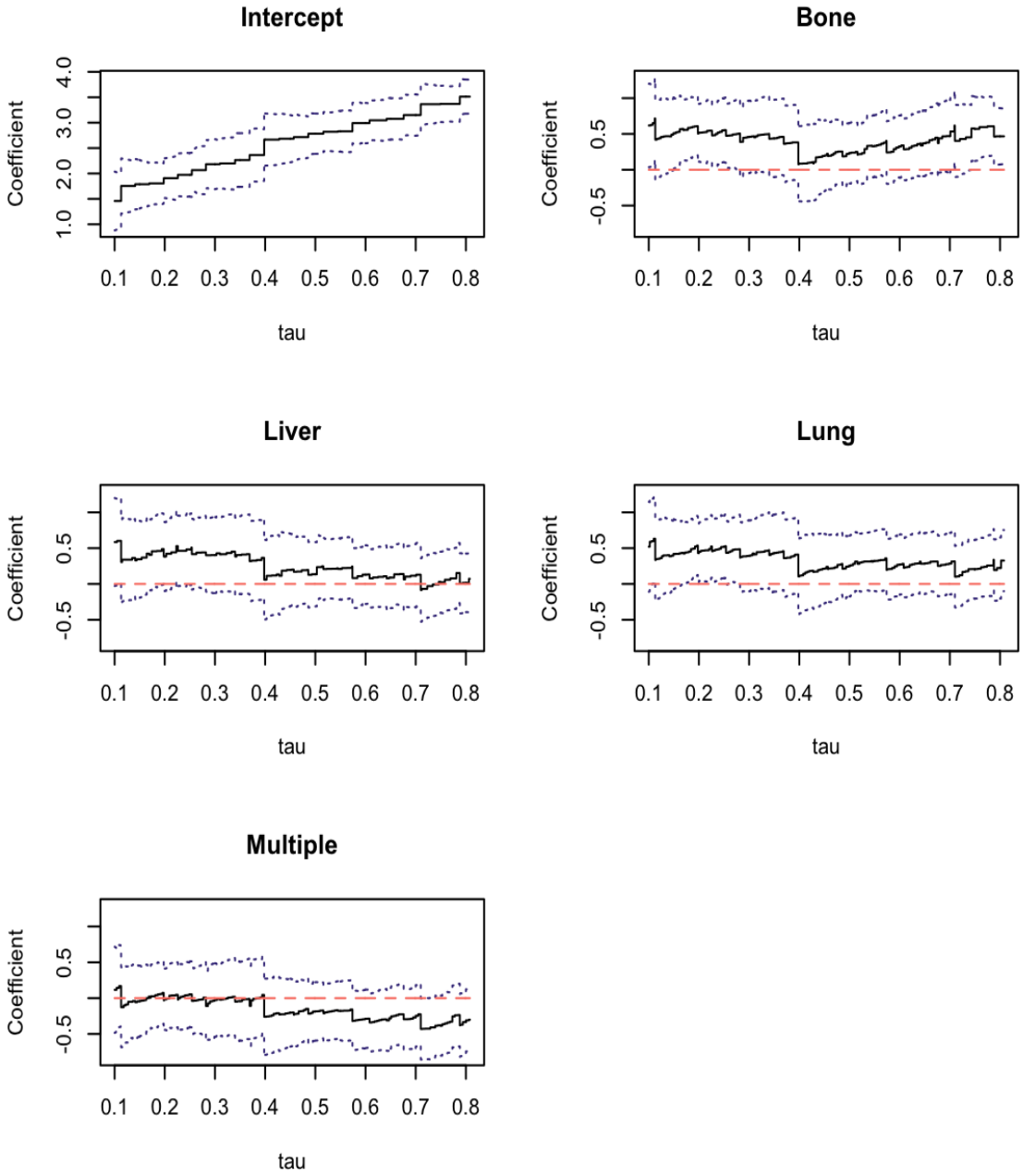


(A) Bone Metastasis

1. $\tau \in [0.1, 0.808]$.

2. Reference category of metastasis covariate: bone.

A.4. TNBC Subtype

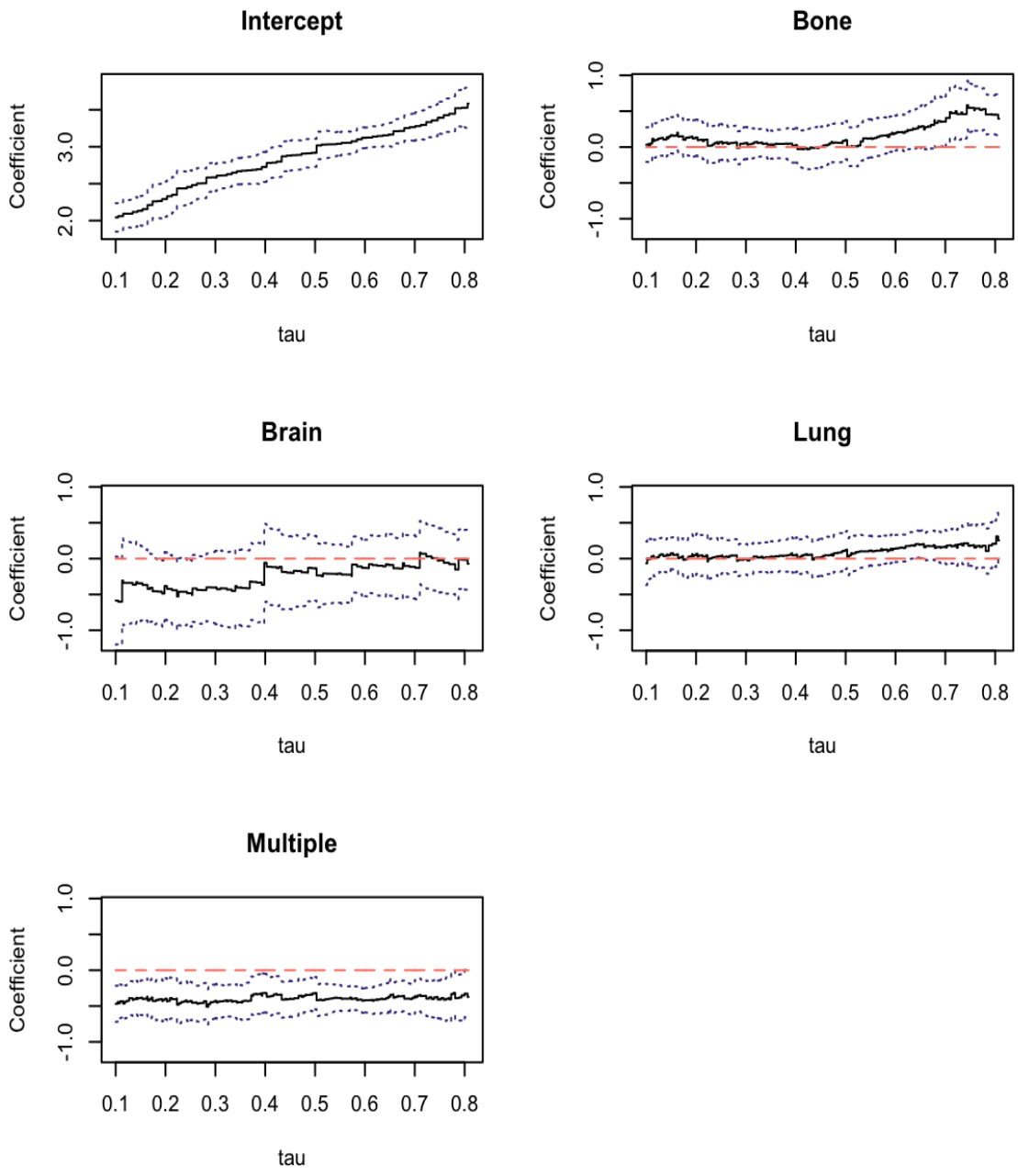


(B) Brain Metastasis

1. $\tau \in [0.1, 0.808]$.

2. Reference category of metastasis covariate: brain.

A.4. TNBC Subtype

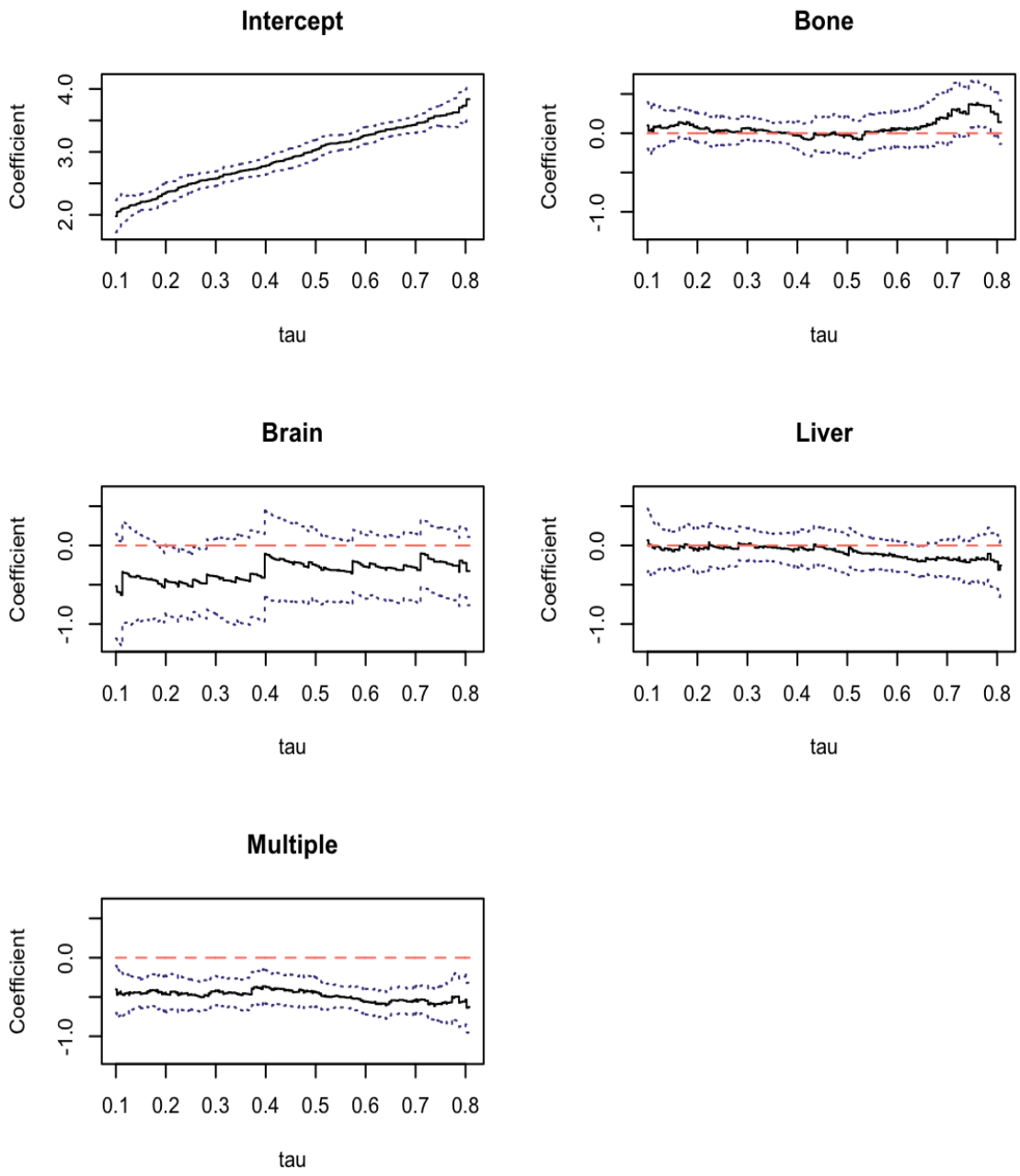


(C) Liver Metastasis

1. $\tau \in [0.1, 0.808]$.

2. Reference category of metastasis covariate: liver.

A.4. TNBC Subtype

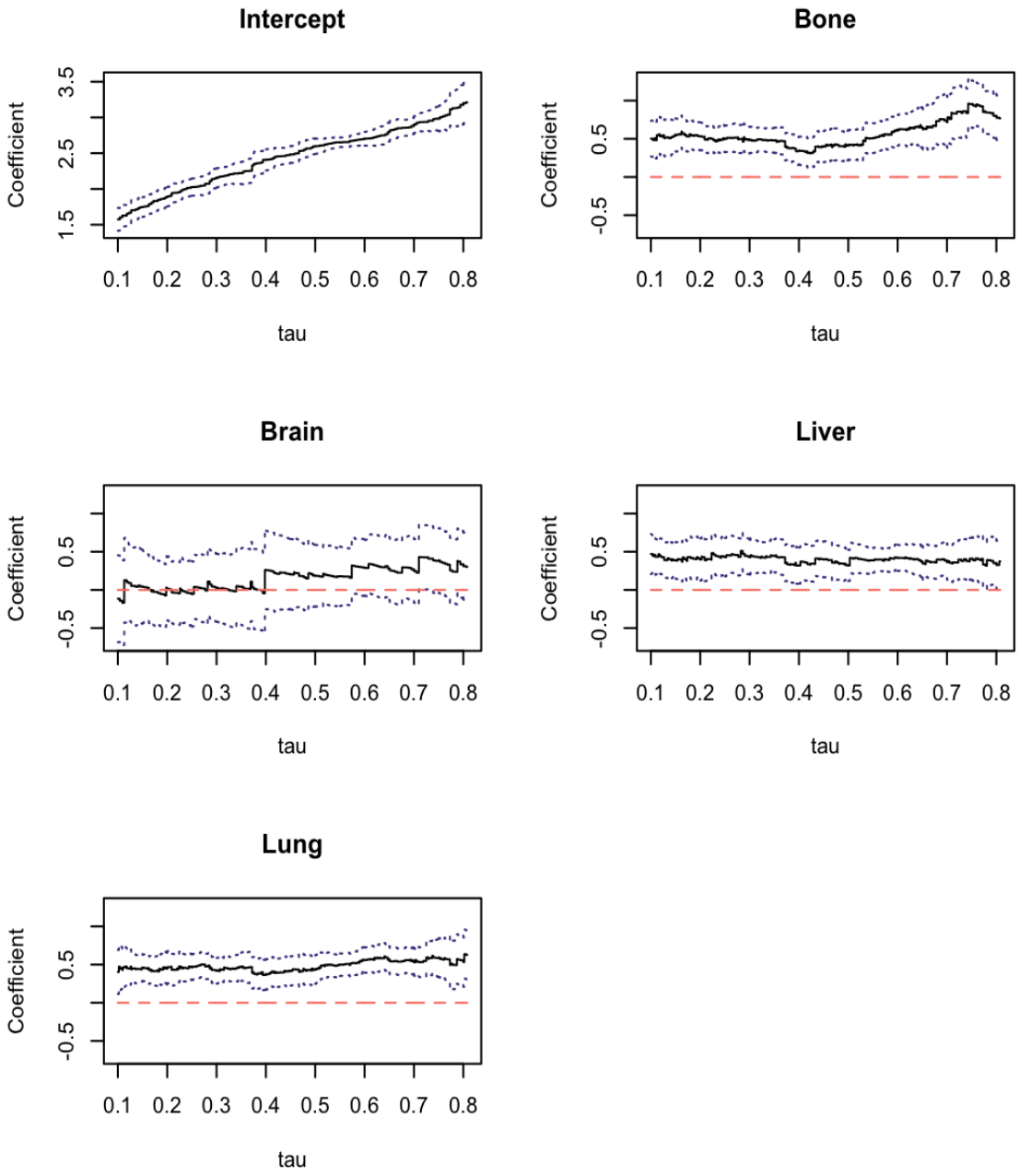


(D) Lung Metastasis

1. $\tau \in [0.1, 0.808]$.

2. Reference category of metastasis covariate: lung.

A.4. TNBC Subtype



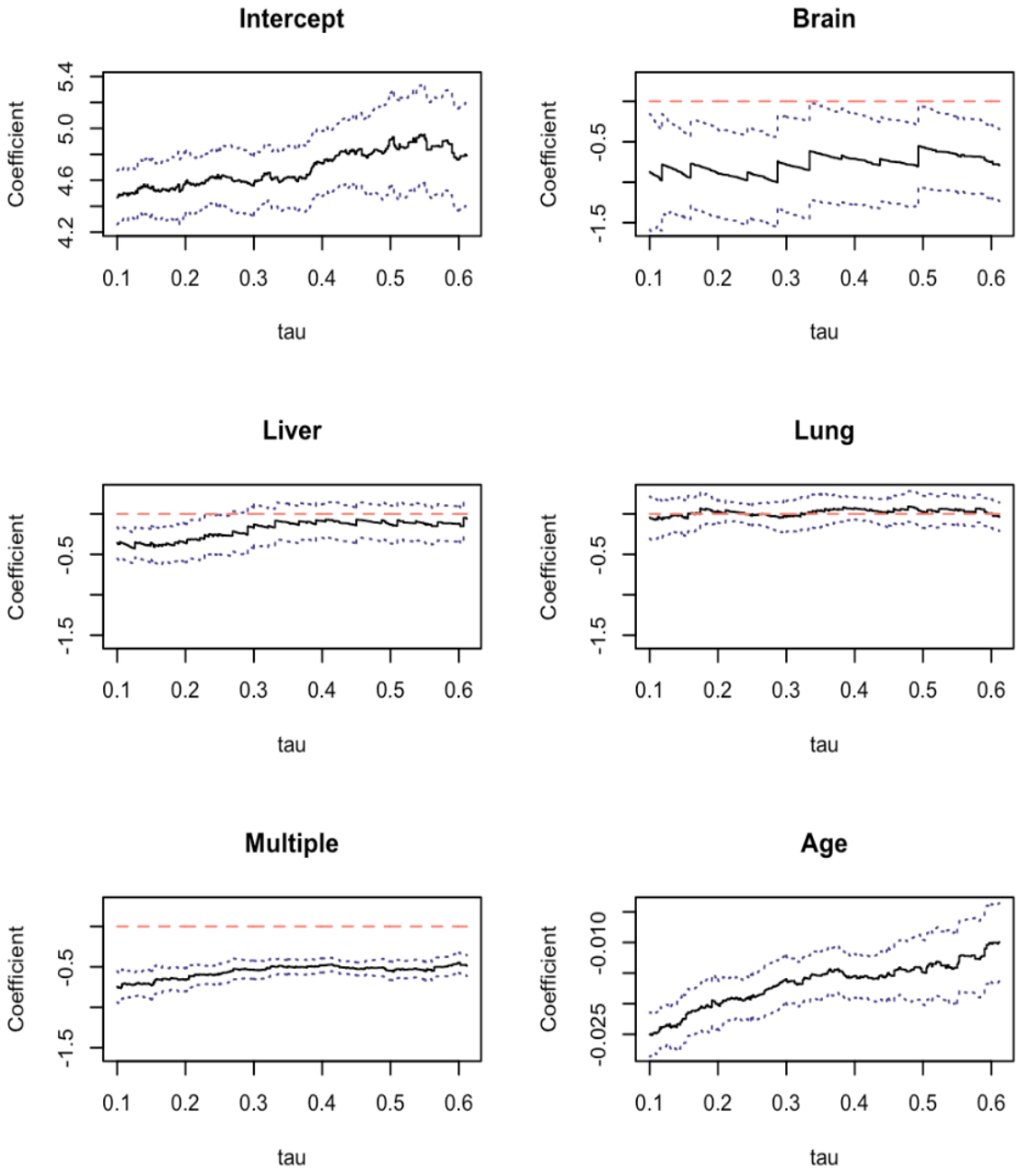
(E) Multiple Metastases

1. $\tau \in [0.1, 0.808]$.

2. Reference category of metastasis covariate: multiple.

Appendix B: Regression Quantiles in Multivariate Analysis

B.1 HR+/HER2- Subtype

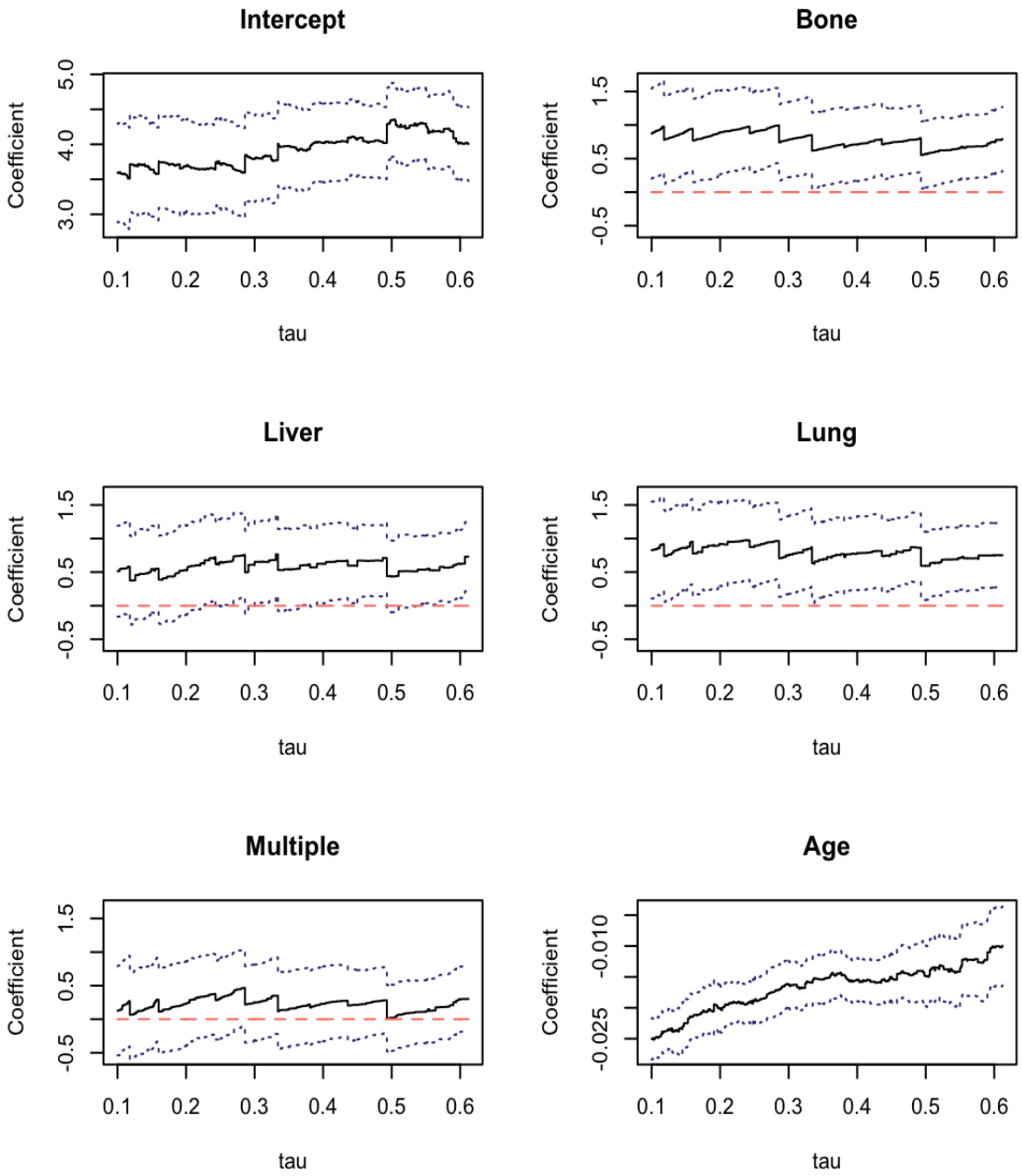


(A) Bone Metastasis

1. $\tau \in [0.1, 0.612]$.

2. Reference category of metastasis covariate: bone.

B.1 HR+/HER2- Subtype

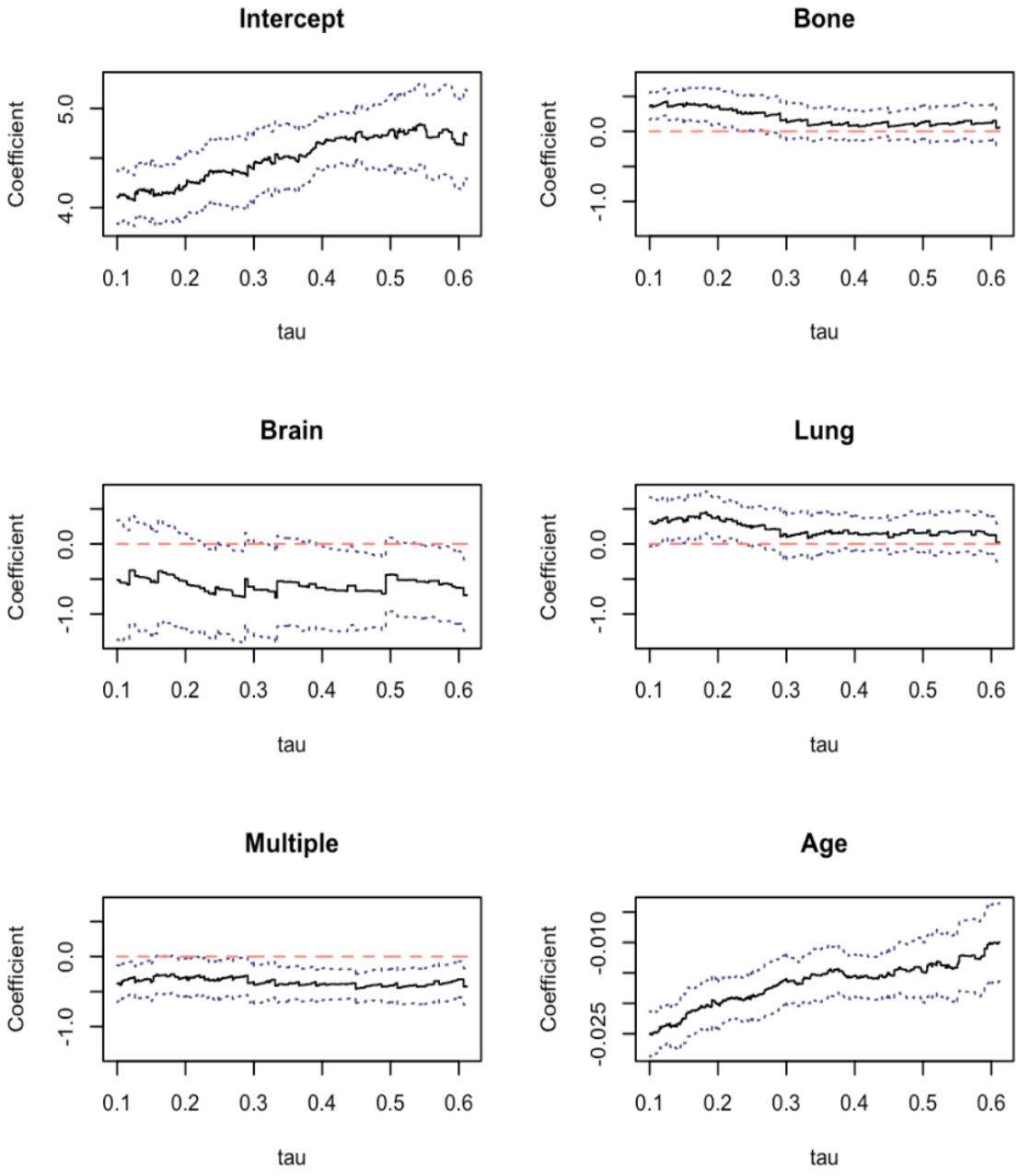


(B) Brain Metastasis

1. $\tau \in [0.1, 0.612]$.

2. Reference category of metastasis covariate: brain.

B.1 HR+/HER2- Subtype

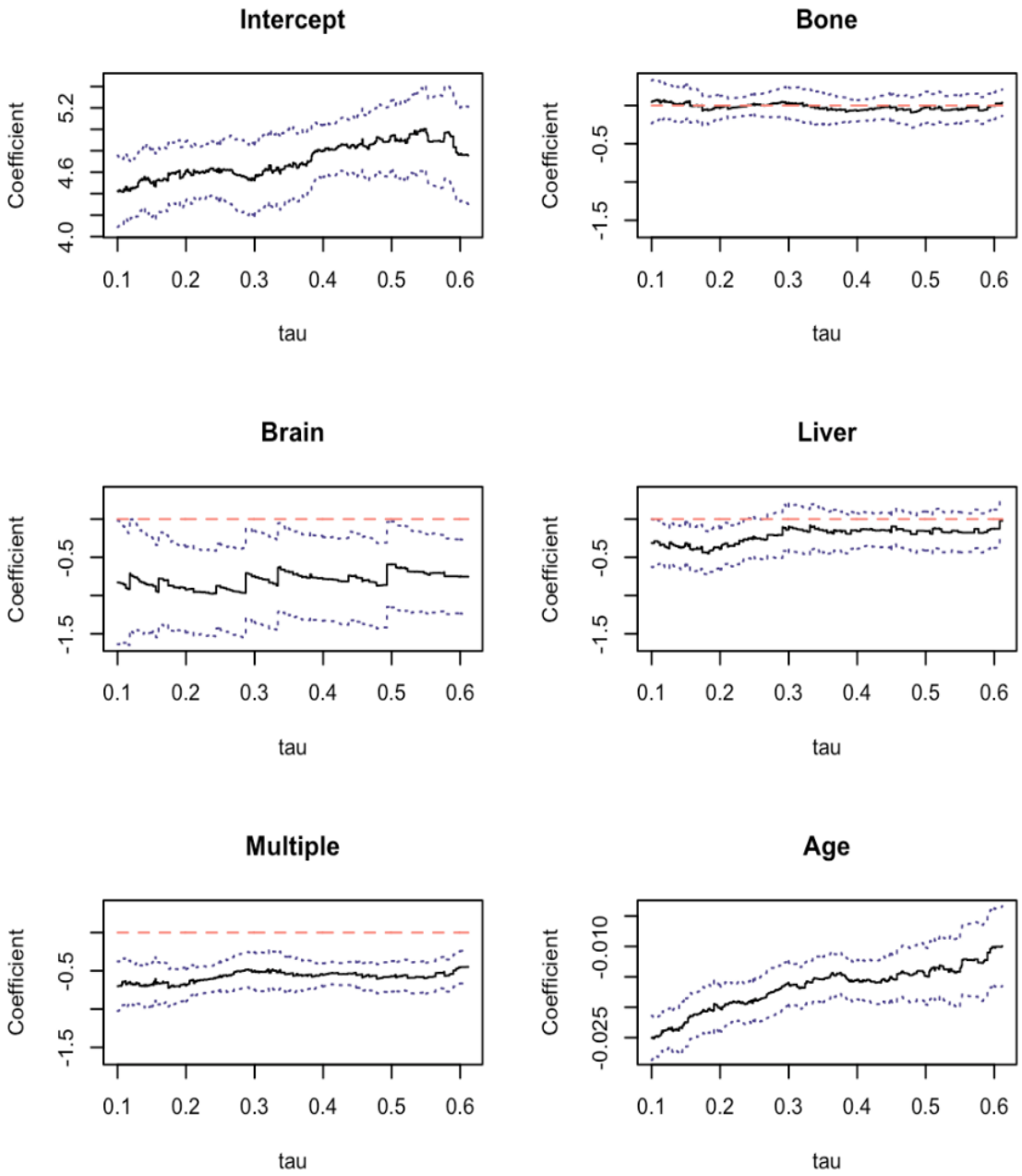


(C) Liver Metastasis

1. $\tau \in [0.1, 0.612]$.

2. Reference category of metastasis covariate: liver.

B.1 HR+/HER2- Subtype

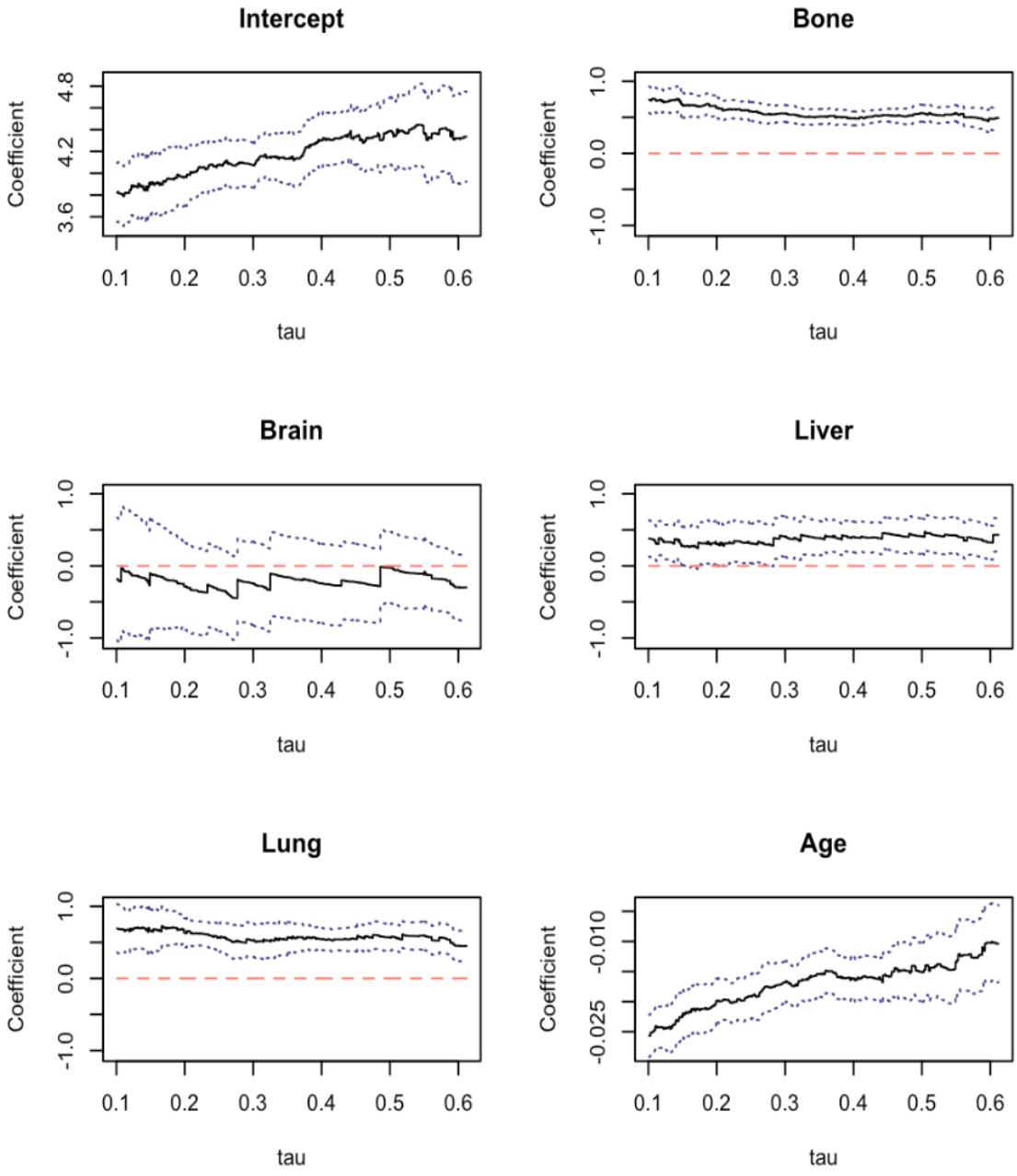


(D) Lung Metastasis

1. $\tau \in [0.1, 0.612]$.

2. Reference category of metastasis covariate: lung.

B.1 HR+/HER2- Subtype

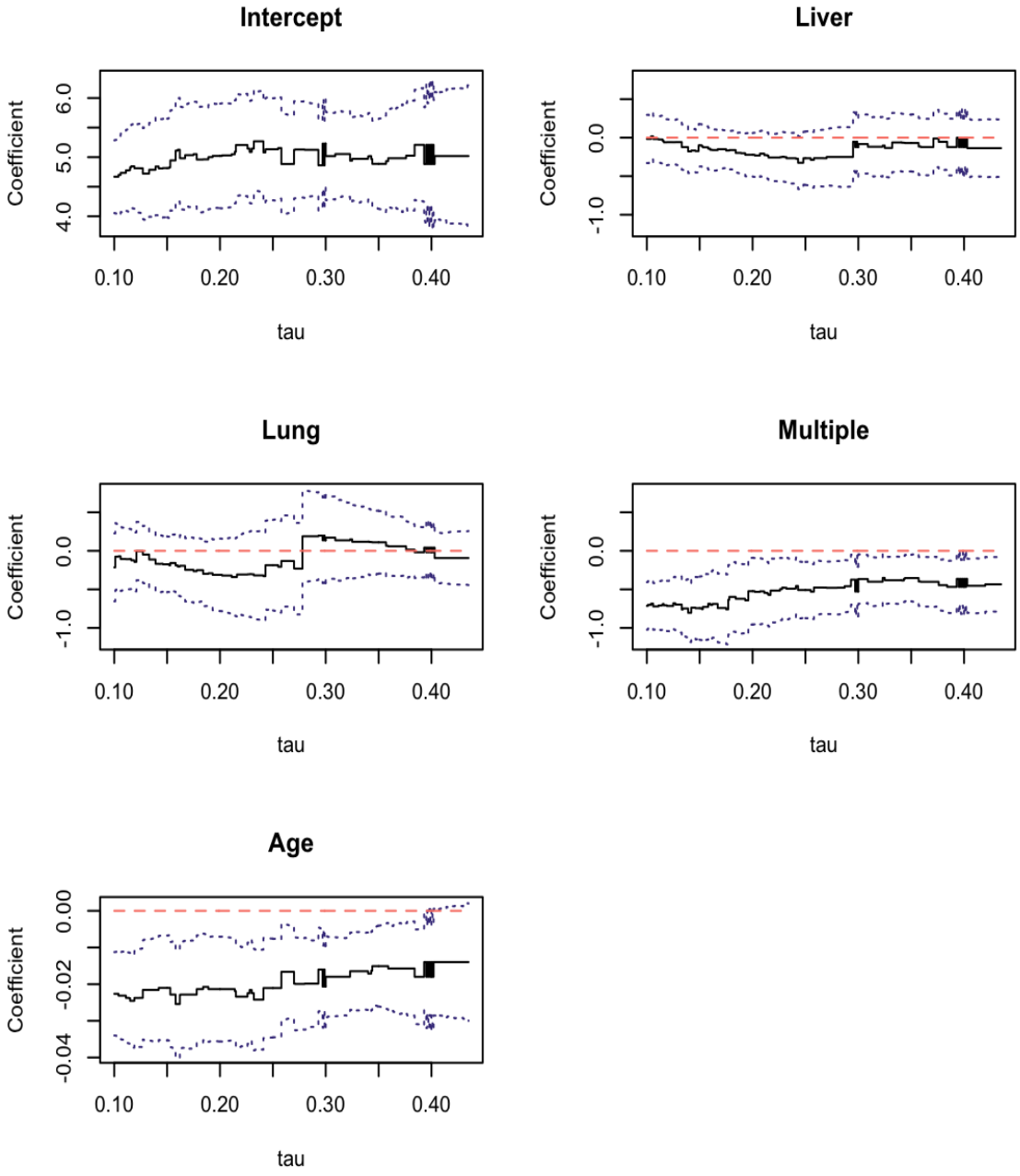


(E) Multiple Metastases

1. $\tau \in [0.1, 0.612]$.

2. Reference category of metastasis covariate: multiple.

B.2 HR+/HER2+ Subtype

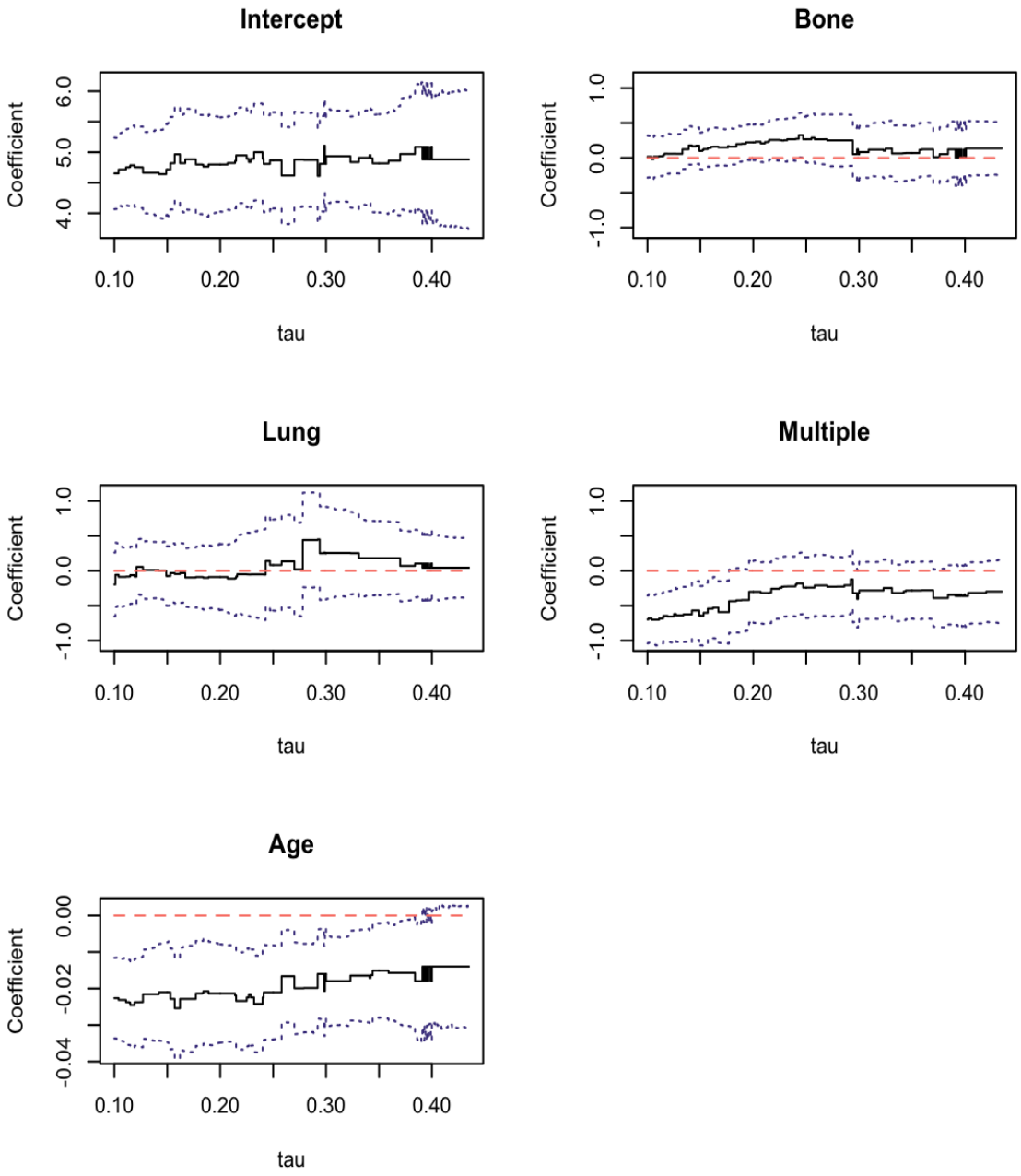


(A) Bone Metastasis

1. $\tau \in [0.1, 0.435]$.

2. Reference category of metastasis covariate: bone.

B.2 HR+/HER2+ Subtype

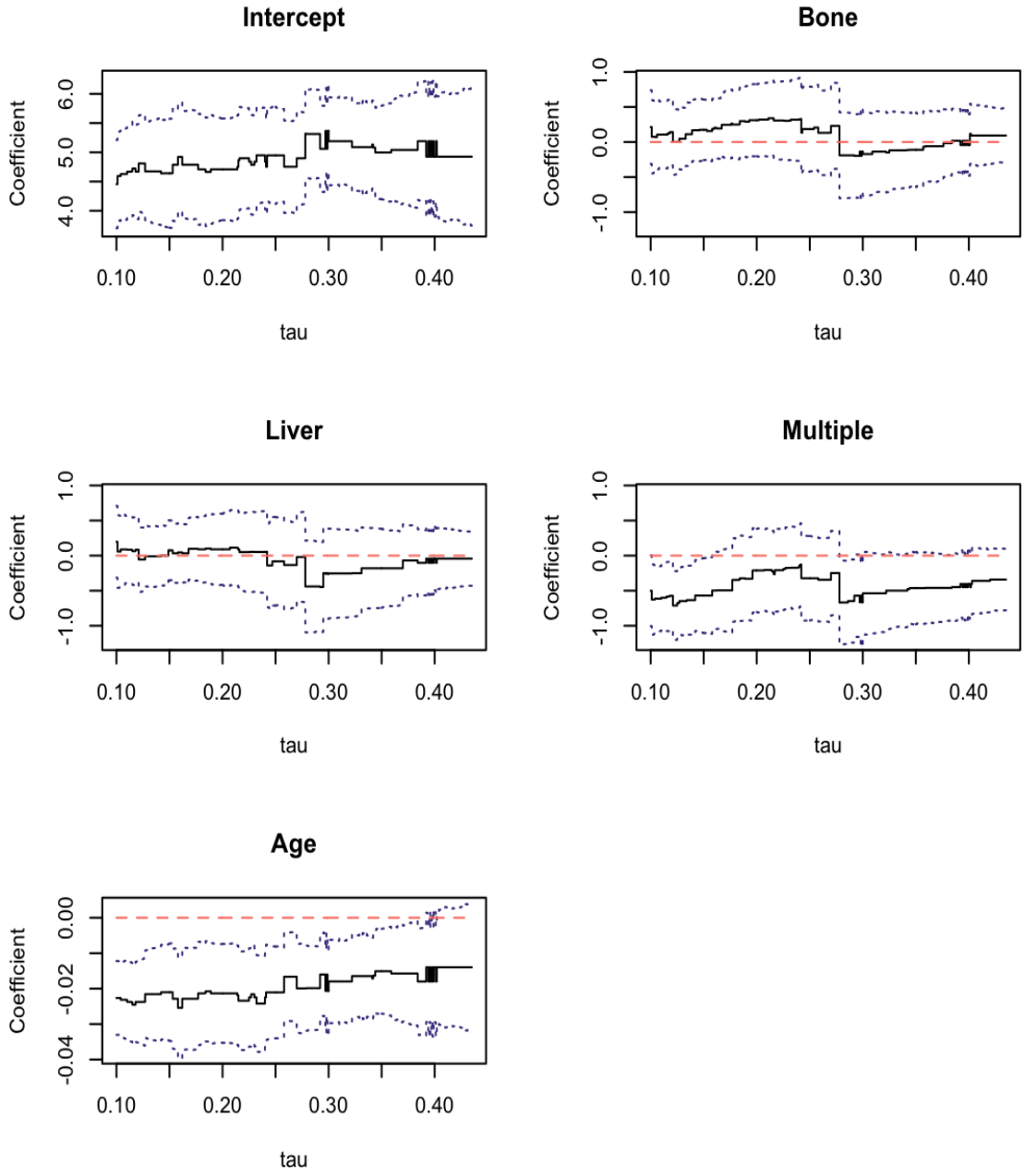


(B) Liver Metastasis

1. $\tau \in [0.1, 0.435]$.

2. Reference category of metastasis covariate: liver.

B.2 HR+/HER2+ Subtype

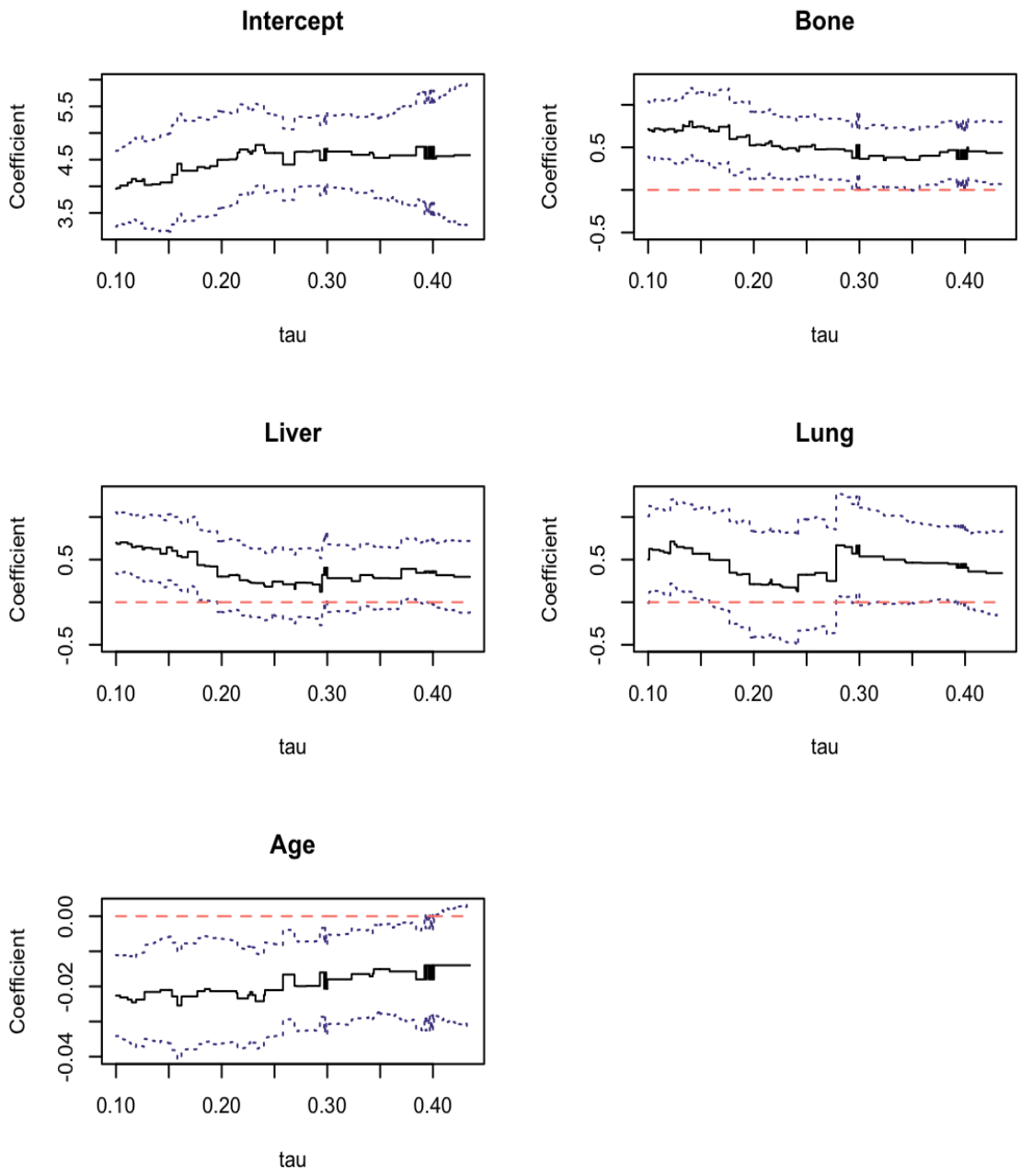


(C) Lung Metastasis

1. $\tau \in [0.1, 0.435]$.

2. Reference category of metastasis covariate: lung.

B.2 HR+/HER2+ Subtype

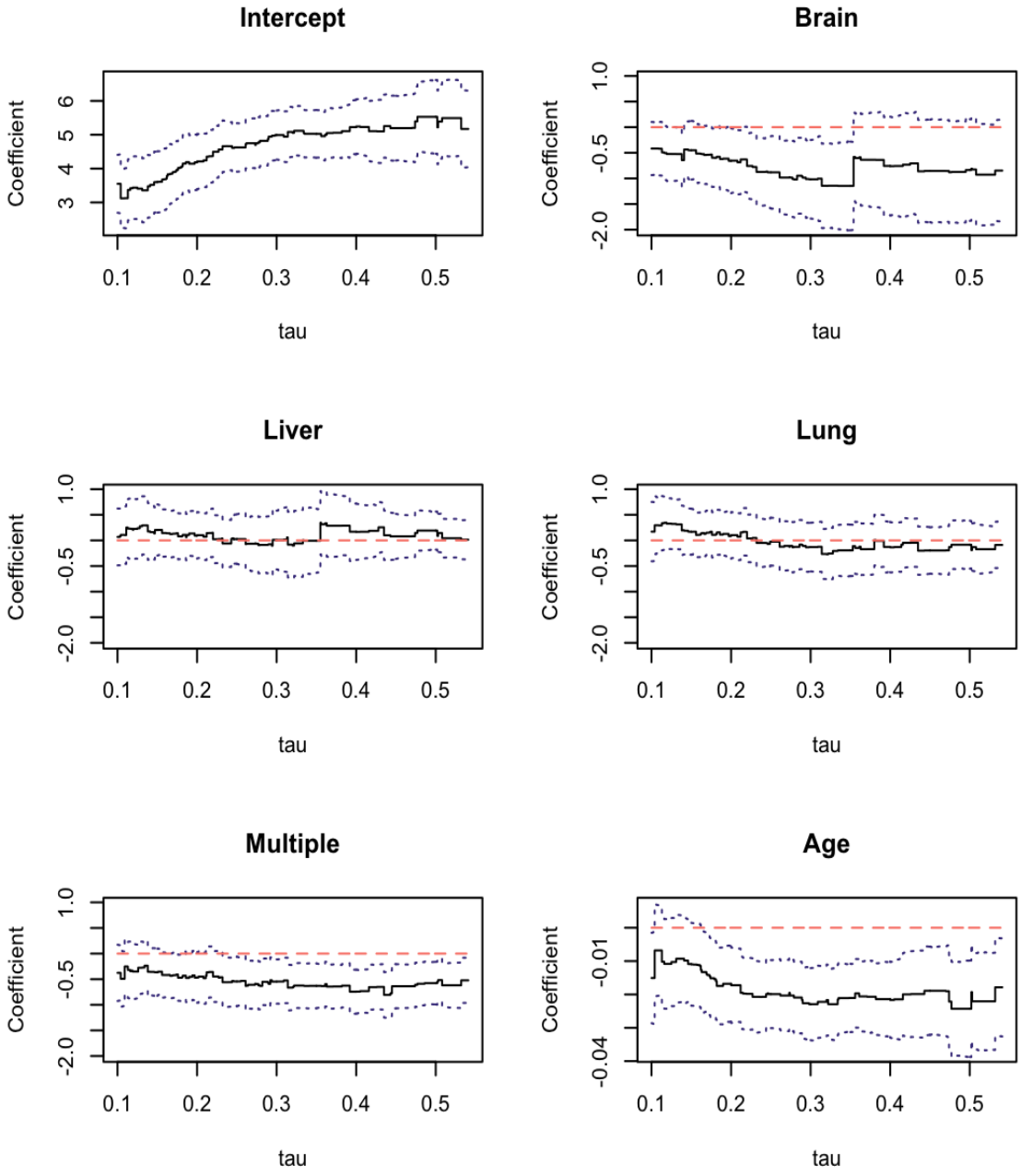


(D) Multiple Metastasis

1. $\tau \in [0.1, 0.435]$.

2. Reference category of metastasis covariate: multiple.

B.3 HR-/HER2+ Subtype

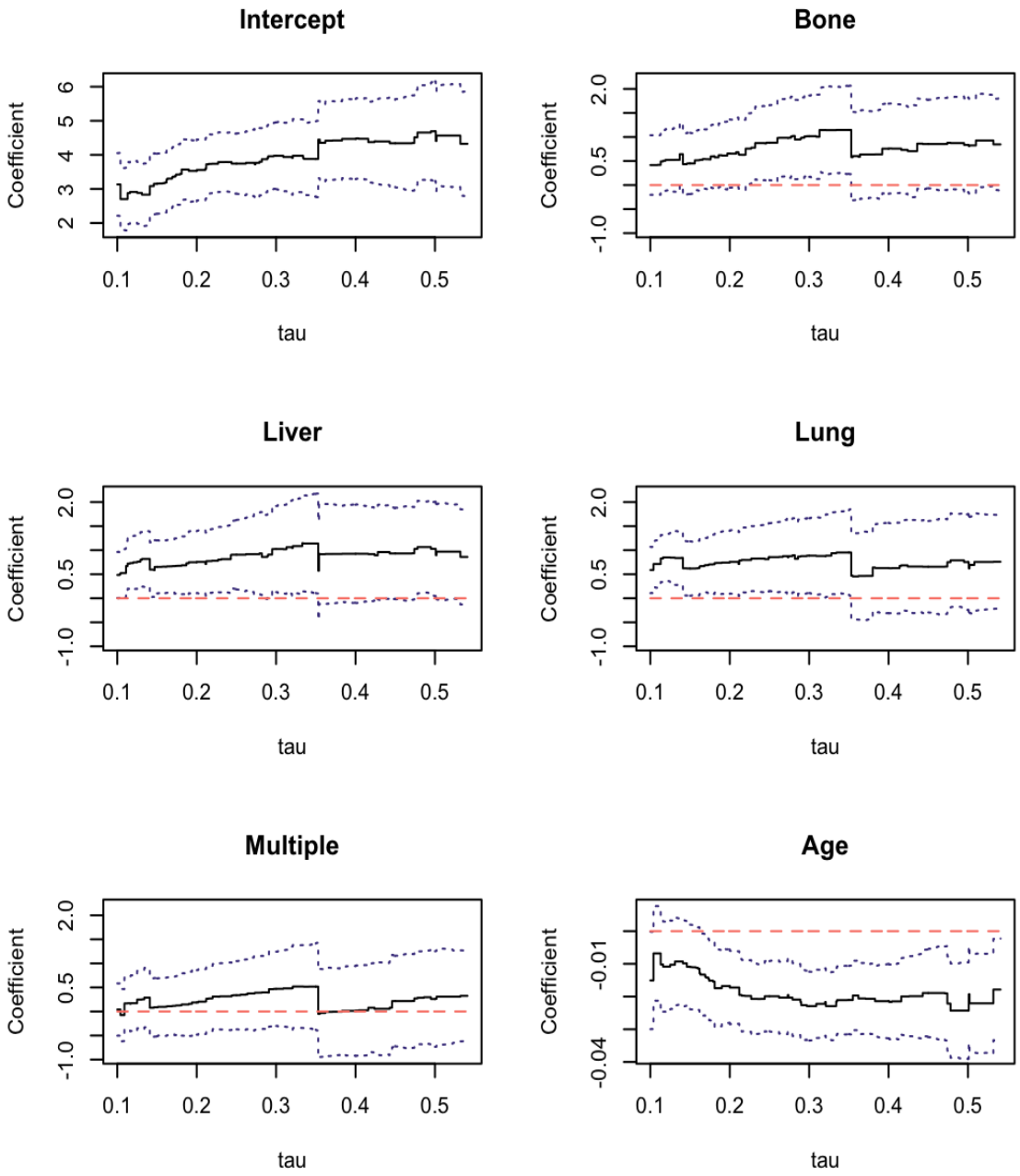


(A) Bone Metastasis

1. $\tau \in [0.1, 0.541]$.

2. Reference category of metastasis covariate: bone.

B.3 HR-/HER2+ Subtype

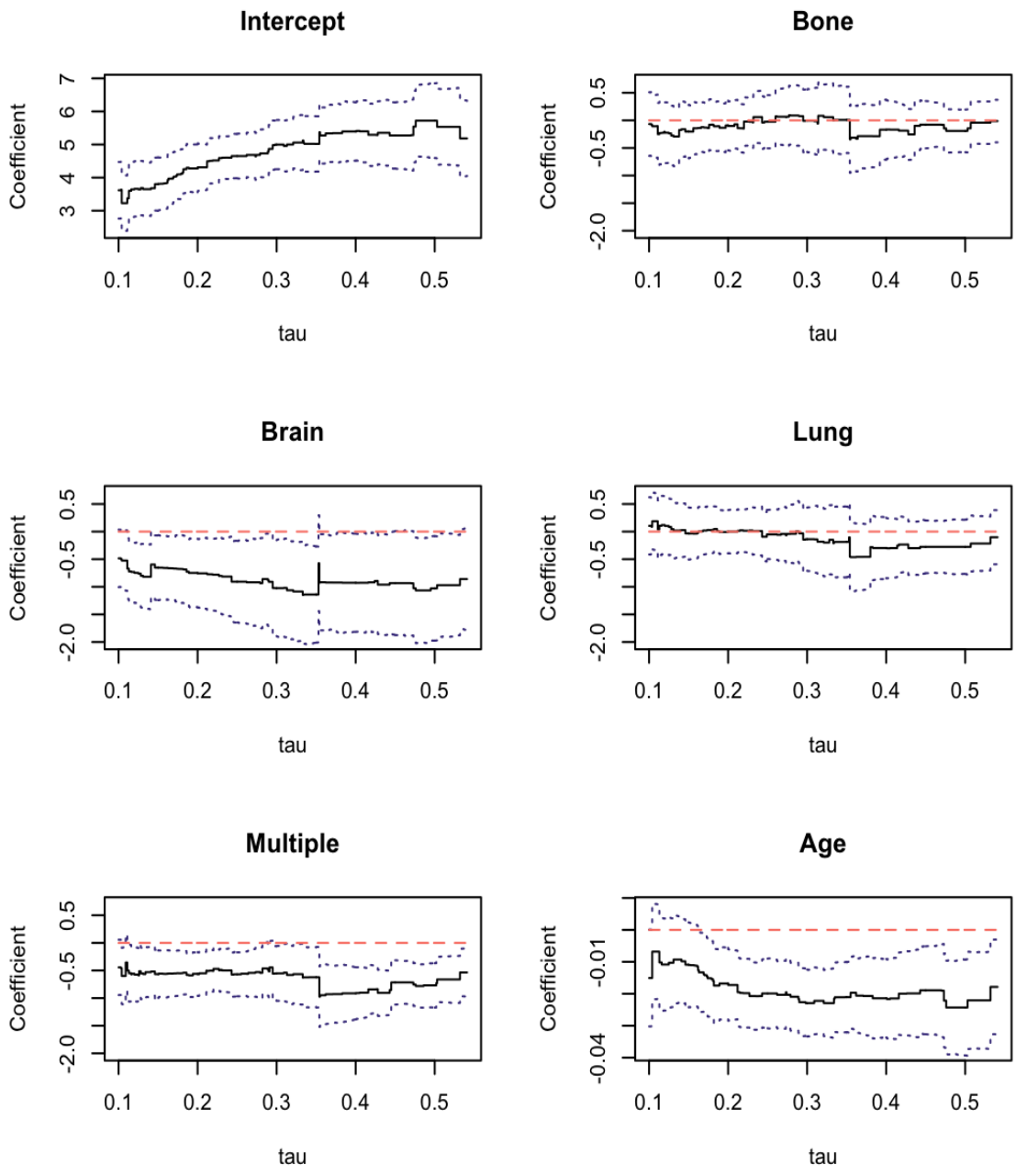


(B) Brain Metastasis

1. $\tau \in [0.1, 0.541]$.

2. Reference category of metastasis covariate: brain.

B.3 HR-/HER2+ Subtype

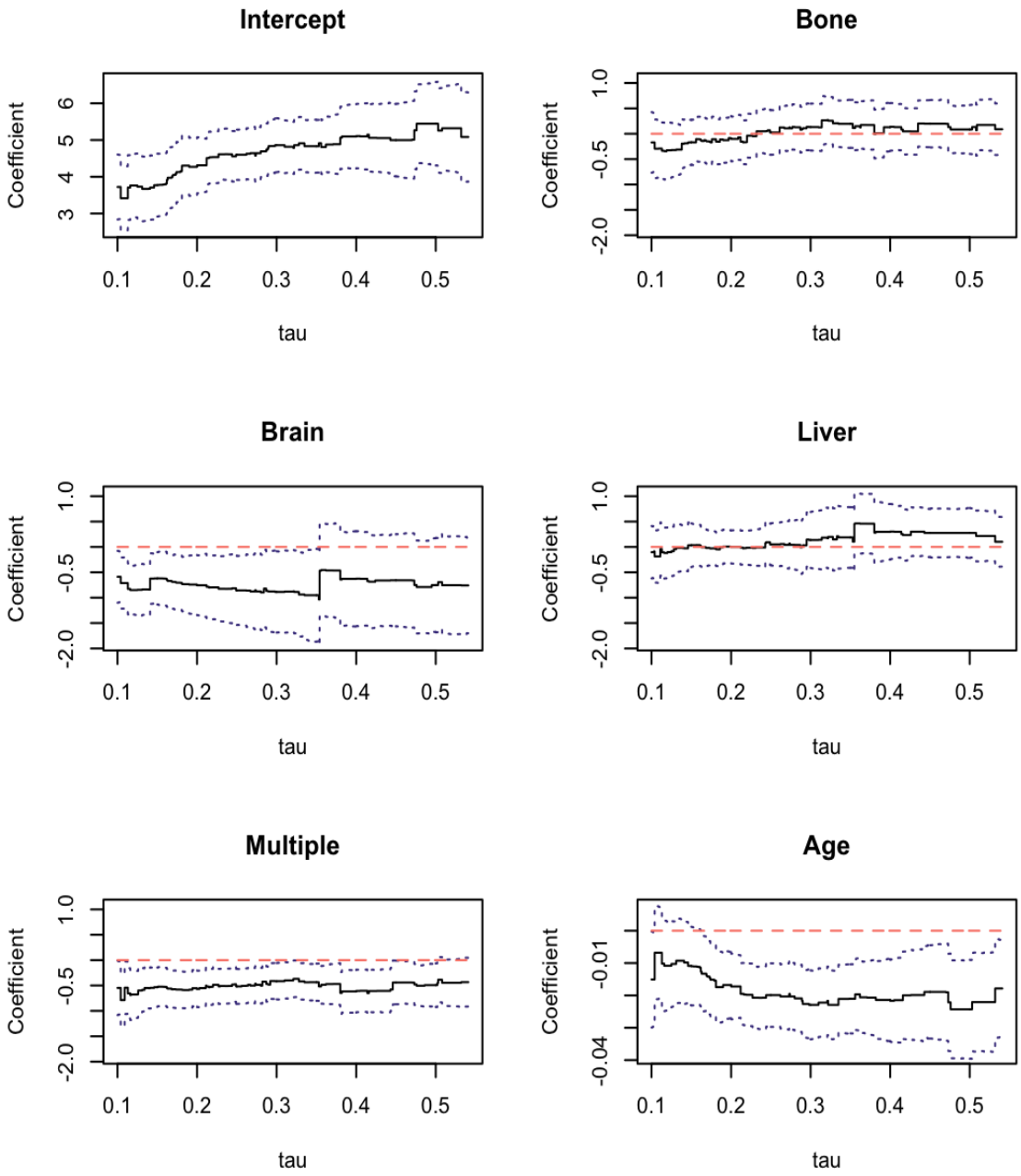


(C) Liver Metastasis

1. $\tau \in [0.1, 0.541]$.

2. Reference category of metastasis covariate: liver.

B.3 HR-/HER2+ Subtype

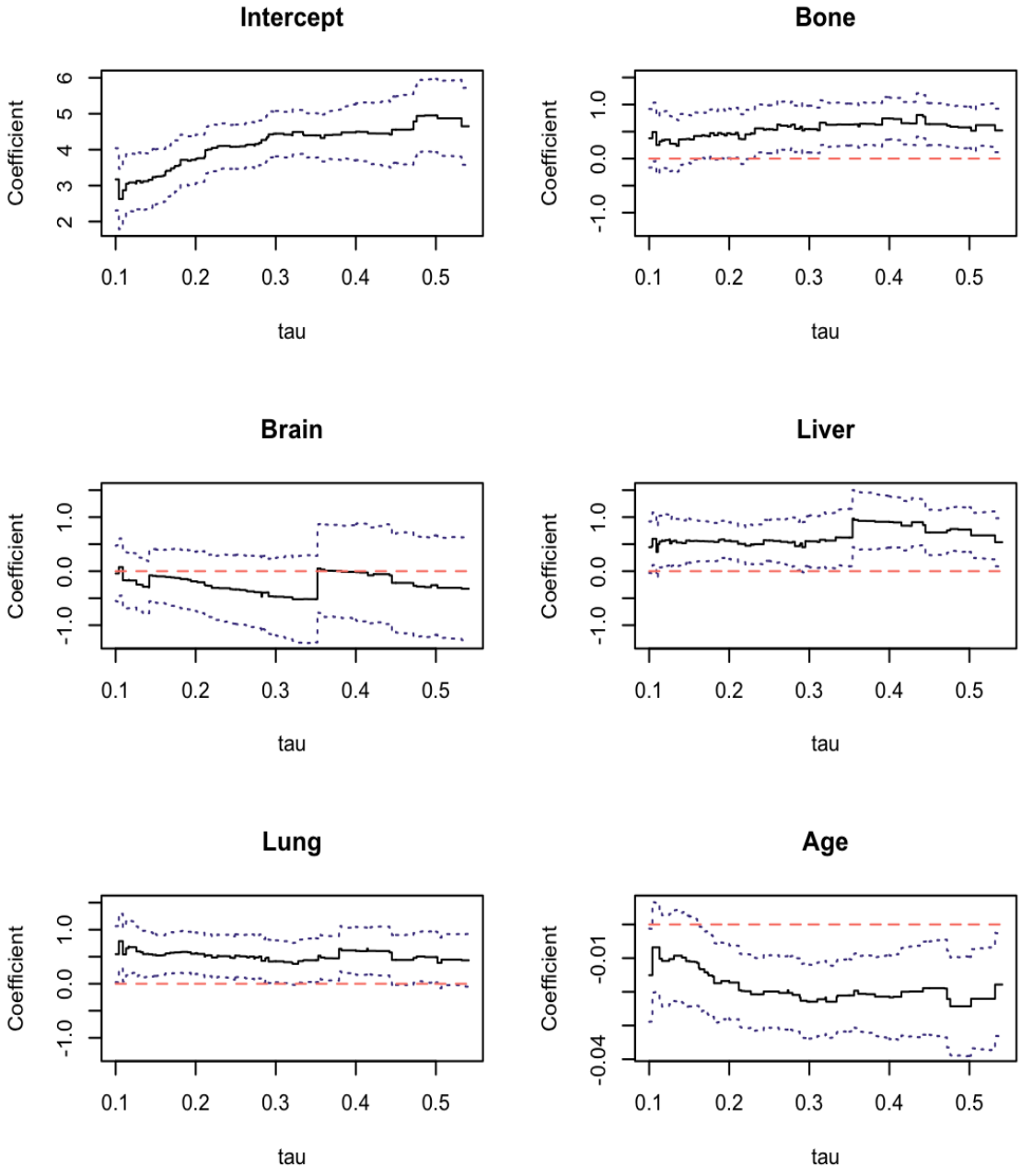


(D) Lung Metastasis

1. $\tau \in [0.1, 0.541]$.

2. Reference category of metastasis covariate: lung.

B.3 HR-/HER2+ Subtype

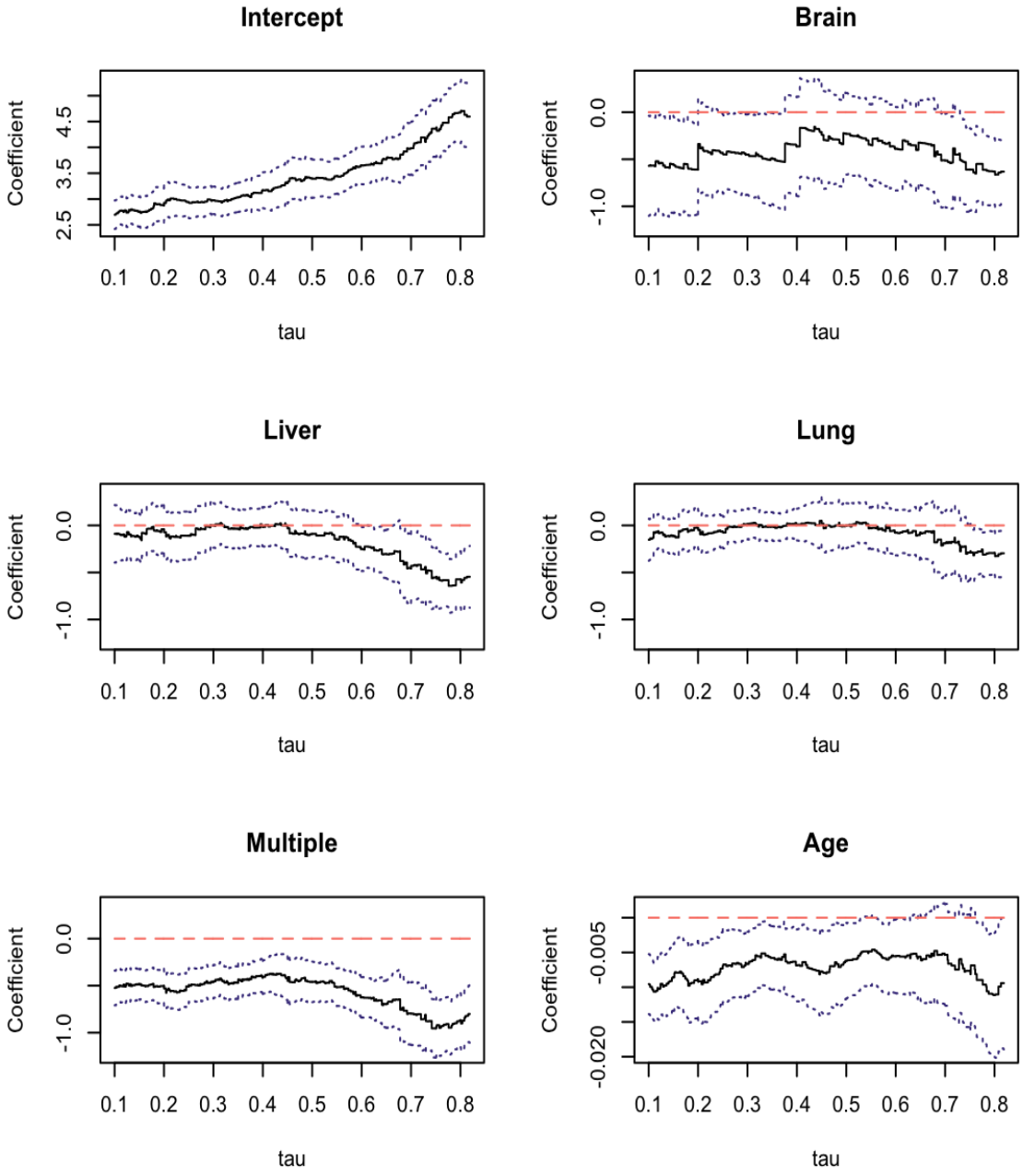


(E) Multiple Metastases

1. $\tau \in [0.1, 0.541]$.

2. Reference category of metastasis covariate: multiple.

B.4 TNBC Subtype

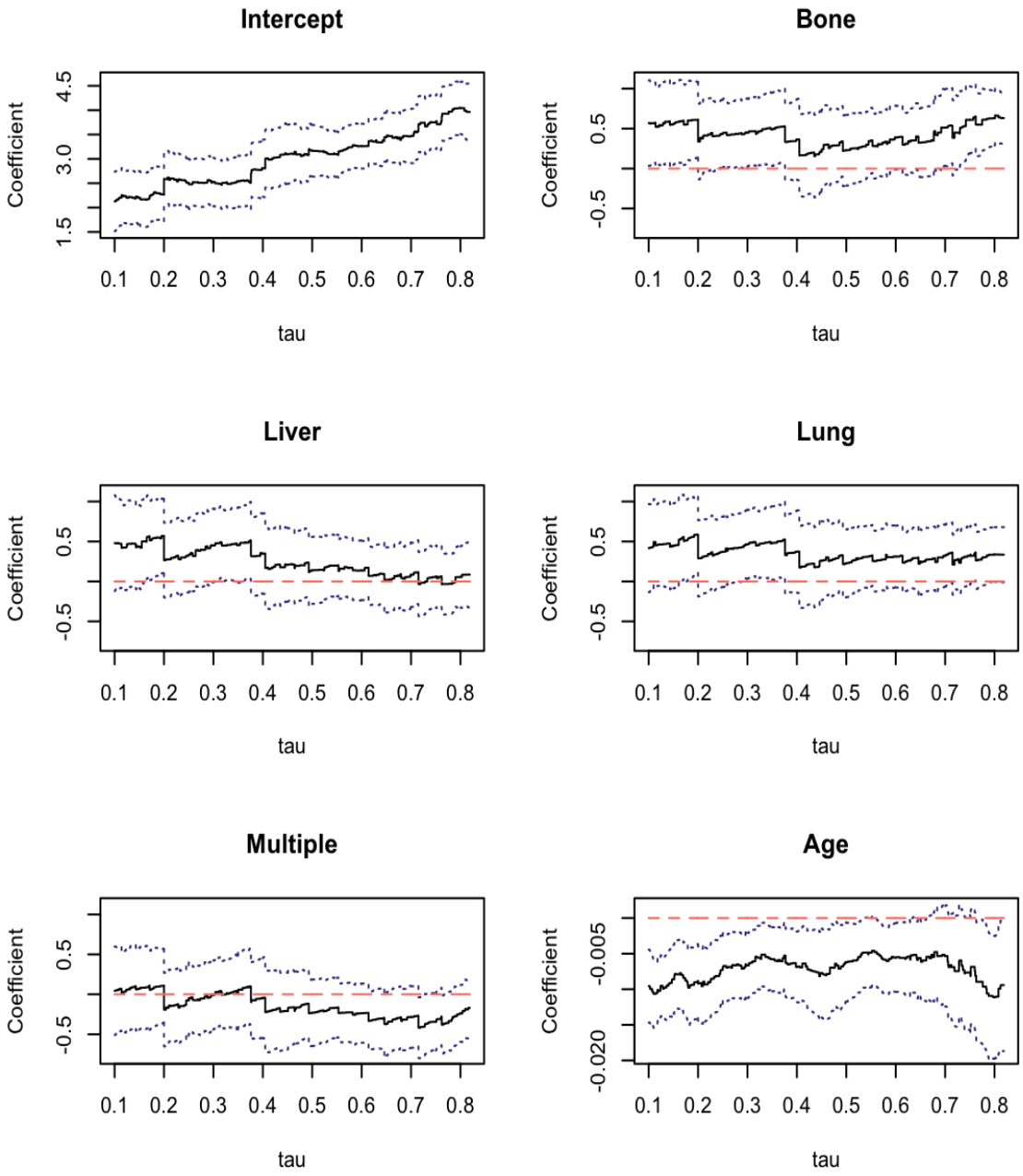


(A) Bone Metastasis

1. $\tau \in [0.1, 0.827]$.

2. Reference category of metastasis covariate: bone.

B.4 TNBC Subtype

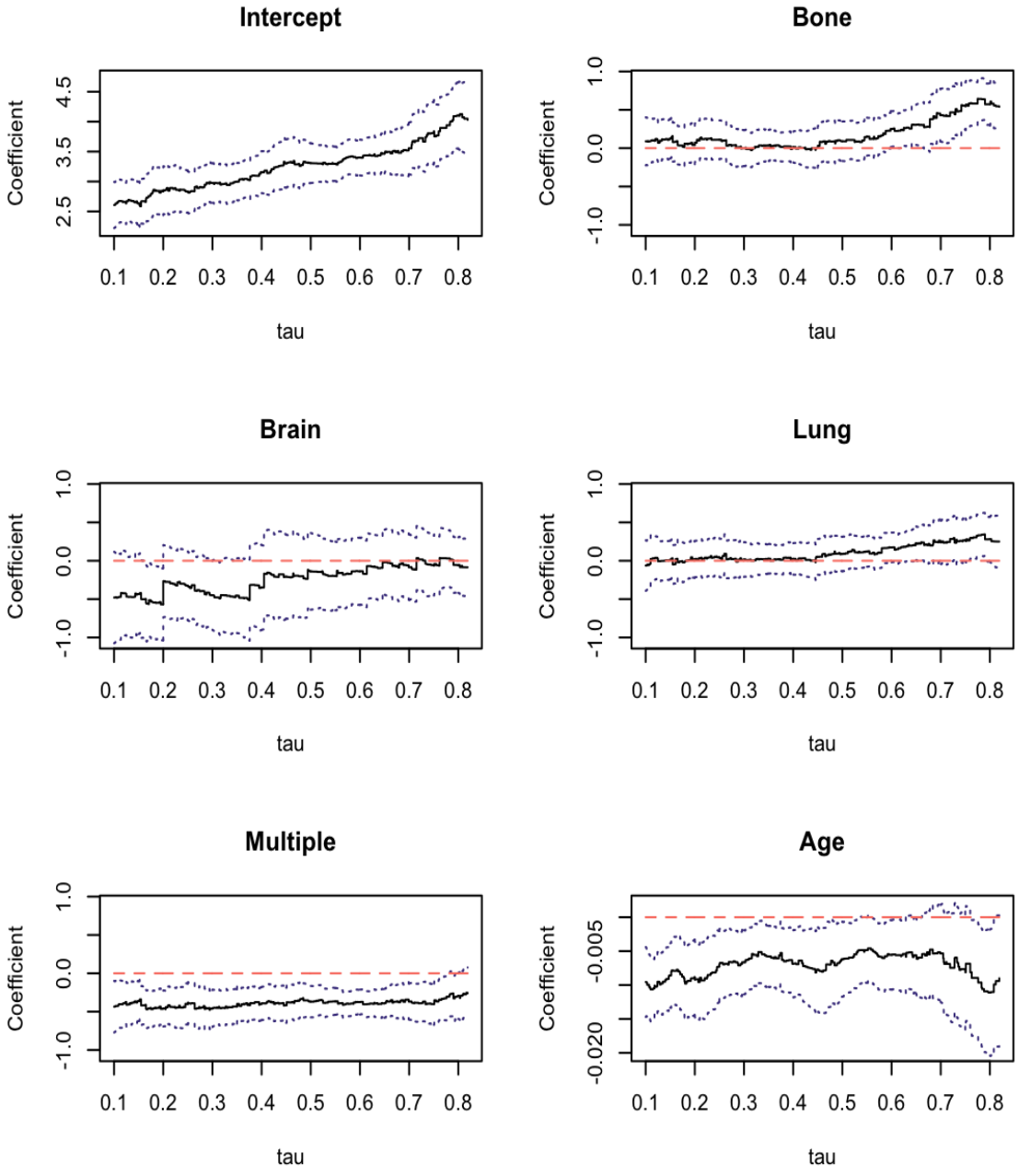


(B) Brain Metastasis

1. $\tau \in [0.1, 0.827]$.

2. Reference category of metastasis covariate: brain.

B.4 TNBC Subtype

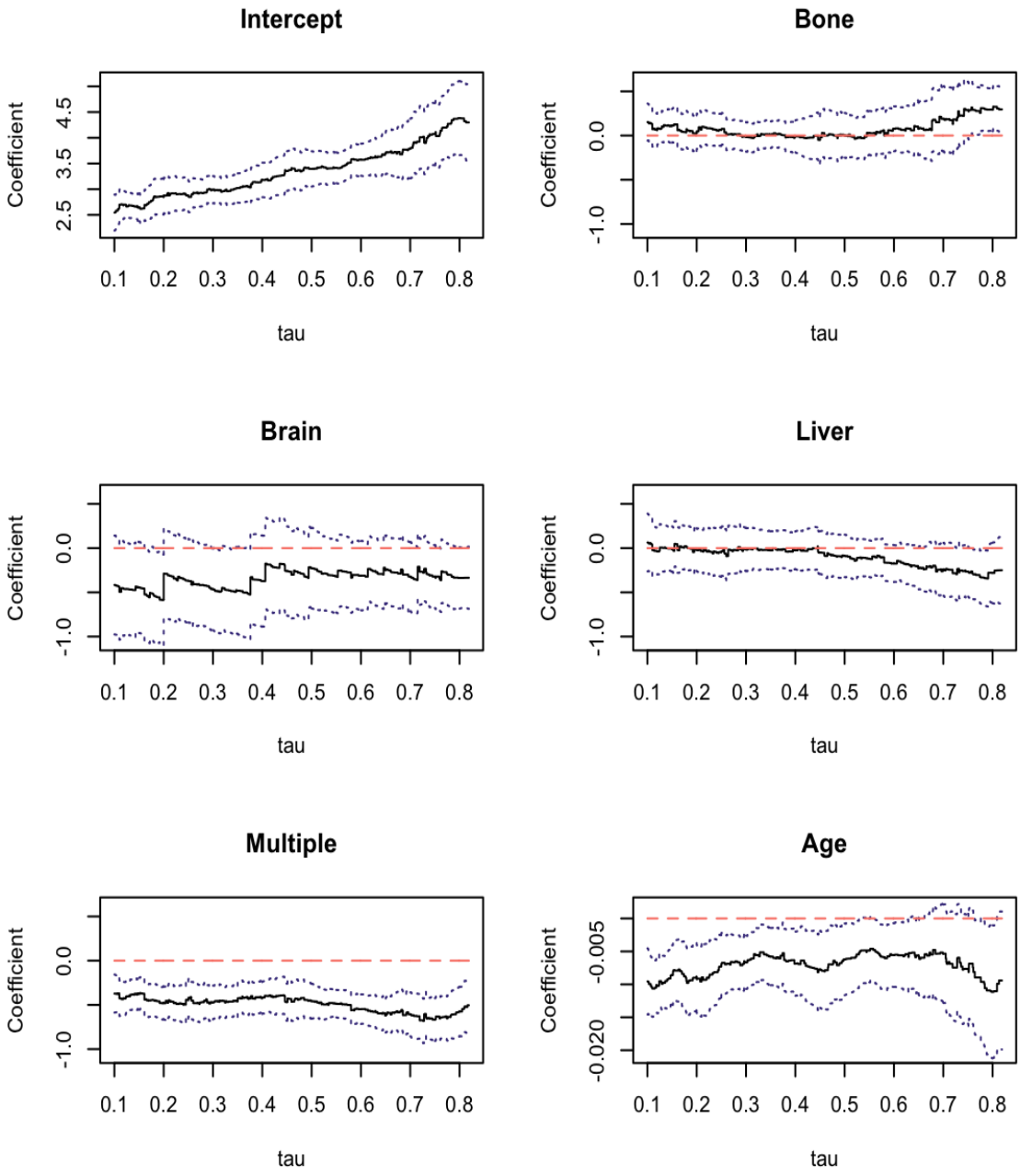


(C) Liver Metastasis

1. $\tau \in [0.1, 0.827]$.

2. Reference category of metastasis covariate: liver.

B.4 TNBC Subtype

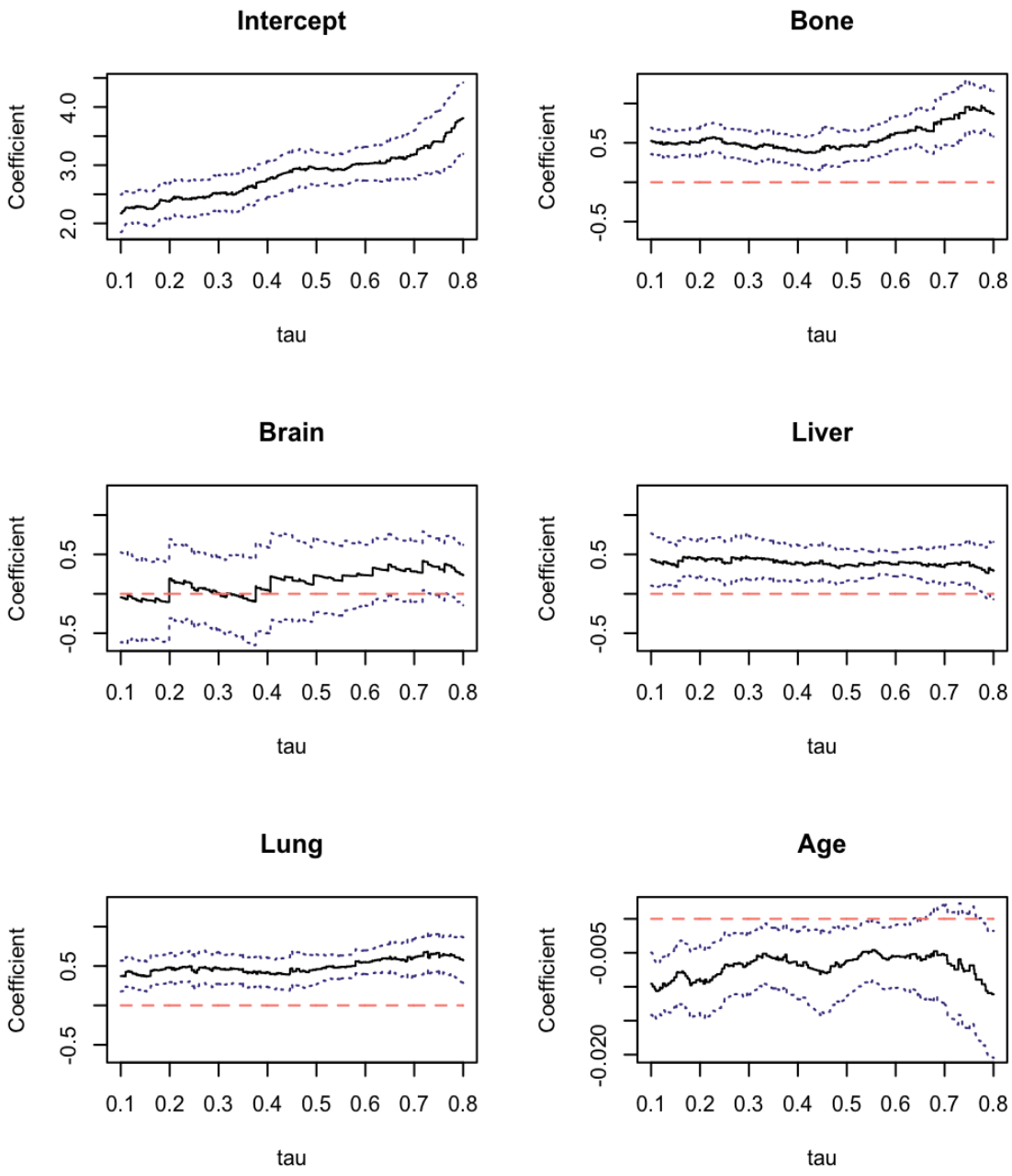


(D) Lung Metastasis

1. $\tau \in [0.1, 0.827]$.

2. Reference category of metastasis covariate: lung.

B.4 TNBC Subtype



(E) Multiple Metastases

1. $\tau \in [0.1, 0.827]$.

2. Reference category of metastasis covariate: multiple.

REFERENCES

1. Siegel RL, Miller KD, Jemal A. Cancer statistics, 2019. CA Cancer J Clin. 2019; 69(1):7-34.

2. Cancer STAT Facts: Female Breast Cancer, Surveillance Epidemiology and End Results Program, 2017

3. Jin X, Mu P. Targeting Breast Cancer Metastasis. Breast Cancer (Auckl), 9(Suppl 1): 23-34, 2015

4. Understanding A Breast Cancer Diagnosis, American Cancer Society, 2017

5. Li X, Yang J, Peng L, et al. Triple-negative breast cancer has worse overall survival and cause-specific survival than non-triple-negative breast cancer. Breast cancer research and treatment. 2017; 161: 279-87.

6. Carey LA, Perou CM, Livasy CA, Dressler LG, Cowan D, Conway K, Karaca G, Troester MA, Tse CK, Edmiston S, Deming SL. Race, breast cancer subtypes, and survival in the Carolina Breast Cancer Study. Jama. 2006 Jun 7;295(21):2492-502.

7. Leone JP, Leone J, Zwenger AO, Iturbe J, Leone BA, Vallejo CT. Prognostic factors and survival according to tumour subtype in women presenting with breast cancer brain metastases at initial diagnosis. European Journal of Cancer. 2017 Mar 1;74:17-25.

8. Anders CK, Deal AM, Miller CR, Khorram C, Meng H, Burrows E, Livasy C, Fritchie K, Ewend MG, Perou CM, Carey LA. The prognostic contribution of clinical breast cancer subtype, age, and race among patients with breast cancer brain metastases. Cancer. 2011 Apr 15;117(8):1602-11.

9. Lamberts SW, Barker WH, Reubi JC, Krenning EP. Somatostatin-receptor imaging in the localization of endocrine tumors. New England Journal of Medicine. 1990 Nov 1;323(18):1246-9.

10. Soerjomataram I, Louwman MW, Ribot JG, Roukema JA, Coebergh JW. An overview of prognostic factors for long-term survivors of breast cancer. Breast cancer research and treatment. 2008 Feb 1;107(3):309-30.

11. Li X, Yang J, Peng L, et al. Triple-negative breast cancer has worse overall survival and cause-specific survival than non-triple-negative breast cancer. Breast cancer research and treatment. 2017; 161: 279-87.

12. Sørlie T, Tibshirani R, Parker J, Hastie T, Marron JS, Nobel A, Deng S, Johnsen H, Pesich R, Geisler S, Demeter J. Repeated observation of breast tumor subtypes in independent gene expression data sets. Proceedings of the national academy of sciences. 2003 Jul 8;100(14):8418-23.

13. Sørlie T, Perou CM, Tibshirani R, Aas T, Geisler S, Johnsen H, Hastie T, Eisen MB, Van De Rijn M, Jeffrey SS, Thorsen T. Gene expression patterns of breast carcinomas distinguish tumor subclasses with clinical implications. Proceedings of the National Academy of Sciences. 2001 Sep 11;98(19):10869-74.

14. Anders CK, Carey LA. Biology, metastatic patterns, and treatment of patients with triple-negative breast cancer. Clinical breast cancer. 2009 Jun 1;9:S73-81.

15. Arciero CA, Guo Y, Jiang R, Behera M, O'Regan R, Peng L, Li X. ER+/HER2+ breast cancer has different metastatic patterns and better survival than ER-/HER2+ breast cancer. Clinical Breast Cancer. 2019 Feb

16. Xiao W, Zheng S, Yang A, Zhang X, Zou Y, Tang H, Xie X. Breast cancer subtypes and the risk of distant metastasis at initial diagnosis: a population-based study. Cancer management and research. 2018;10:5329.

17. Gong Y, Liu YR, Ji P, Hu X, Shao ZM. Impact of molecular subtypes on metastatic breast cancer patients: a SEER population-based study. Scientific reports. 2017 Mar 27;7:45411.

18. Ellsworth RE, Blackburn HL, Shriver CD, Soon-Shiong P, Ellsworth DL. Molecular heterogeneity in breast cancer: state of the science and implications for patient care. InSeminars in cell & developmental biology 2017 Apr 1 (Vol. 64, pp. 65-72). Academic Press.

19. Turashvili G, Brogi E. Tumor Heterogeneity in Breast Cancer. Frontiers in Medicine. 2017 Dec 8; 4:227.

20. Adjuvant Therapy for Breast Cancer [Internet]. Memorial Sloan Kettering Cancer Center. [cited 2019Mar19]. Available from: https://www.mskcc.org/cancer-care/patient-education/adjuvant-therapy-breast

21. Chew HK. Adjuvant therapy for breast cancer: who should get what?. Western Journal of medicine. 2001 Apr;174(4):284.

22. Adjuvant Therapy for Breast Cancer [Internet]. Memorial Sloan Kettering Cancer Center. [cited 2019Mar19]. Available from: https://www.mskcc.org/cancer-care/patient-education/adjuvant-therapy-breast

23. Swan J, Wingo P, Clive R, West D, Miller D, Hutchison C, Sondik EJ, Edwards BK. Cancer surveillance in the US: can we have a national system?. Cancer: Interdisciplinary International Journal of the American Cancer Society. 1998 Oct 1;83(7):1282-91.

24. Jagsi, R., Bekelman, J. E., Chen, A., Chen, R. C., Hoffman, K., Shih, Y. C. T., ... & James, B. Y. (2014). Considerations for observational research using large data sets in radiation oncology. International Journal of Radiation Oncology* Biology* Physics, 90(1), 11-24.

25. SEER Training Modules, Cancer Registration & Surveillance Modules. U. S. National Institutes of Health, National Cancer Institute. 03 Feb. 2019

<https://training.seer.cancer.gov/registration/types/hospital.html/>.

26. Merkow, R. P., Rademaker, A. W., & Bilimoria, K. Y. (2018). Practical Guide to Surgical Data Sets: National Cancer Database (NCDB). JAMA surgery.

27. Zahnd WE, Jenkins WD, James AS, Izadi SR, Steward DE, Fogleman AJ, Colditz GA, Brard L. Utility and Generalizability of Multi-State, Population-Based Cancer Registry Data for Rural Cancer Surveillance Research in the United States. Cancer Epidemiology and Prevention Biomarkers. 2018 Jan 1:cebp-1087.

28. Feig B. Comprehensive databases: A cautionary note. Annals of surgical oncology. 2013 Jun 1;20(6):1756-8.

29. James BY, Gross CP, Wilson LD, Smith BD. NCI SEER public-use data: applications and limitations in oncology research. Oncology. 2009 Mar 1;23(3):288-288.

30. Arnold BN, Thomas DC, Rosen JE, Salazar MC, Blasberg JD, Boffa DJ, Detterbeck FC, Kim AW. Lung cancer in the very young: treatment and survival in the national cancer data base. Journal of Thoracic Oncology. 2016 Jul 1;11(7):1121-31.

31. Noone AM, Lund JL, Mariotto A, Cronin K, McNeel T, Deapen D, Warren JL. Comparison of SEER treatment data with Medicare claims. Medical care. 2016 Sep;54(9):e55.

32. Zhang Z. Semi-parametric regression model for survival data: graphical visualization with R. Annals of translational medicine. 2016 Dec;4(23).

33. Harrell FE. Cox proportional hazards regression model. InRegression modeling strategies 2015 (pp. 475-519). Springer, Cham.

34. Abreu MH, Afonso N, Abreu PH, Menezes F, Lopes P, Henrique R, Pereira D, Lopes C. Male breast cancer: Looking for better prognostic subgroups. The Breast. 2016 Apr 1;26:18-24.

35. Fisher S, Gao H, Yasui Y, Dabbs K, Winget M. Survival in stage I–III breast cancer patients by surgical treatment in a publicly funded health care system. Annals of Oncology. 2015 Feb 23;26(6):1161-9.

36. Tichy JR, Deal AM, Anders CK, Reeder-Hayes K, Carey LA. Race, response to chemotherapy, and outcome within clinical breast cancer subtypes. Breast cancer research and treatment. 2015 Apr 1;150(3):667-74.

37. Gong Q, Schaubel DE. Estimating the average treatment effect on survival based on observational data and using partly conditional modeling. Biometrics. 2017 Mar 1;73(1):134-44.

38. Walters SJ. What is a Cox model?. Newmarket, England: Hayward Medical Communications; 1999 Mar.

39. Portnoy S. Censored regression quantiles. Journal of the American Statistical Association. 2003 Dec 1;98(464):1001-12.

40. Gorfine M, Goldberg Y, Ritov YA. A quantile regression model for failure-time data with time-dependent covariates. Biostatistics. 2017 Jan 1;18(1):132-46.

41. Qian J, Peng L. Censored quantile regression with partially functional effects. Biometrika. 2010 Oct 15;97(4):839-50.

42. Koenker, R. (2011). Censored Quantile Regression and Survival Models.

43. Wu Y, Ma Y, Yin G. Smoothed and corrected score approach to censored quantile regression with measurement errors. Journal of the American Statistical Association. 2015 Oct 2;110(512):1670-83.

44. Cox DR. Partial likelihood. Biometrika. 1975 Aug 1;62(2):269-76.

45. Simon N, Friedman J, Hastie T, Tibshirani R. Regularization paths for Cox's proportional hazards model via coordinate descent. Journal of statistical software. 2011 Mar;39(5):1.

46. Breslow NE. Contribution to discussion of paper by DR Cox. J. Roy. Statist. Soc., Ser. B. 1972;34:216-7.

47. Breheny P. Tied survival times; estimation of survival probabilities.

48. Leng C, Tong X. A quantile regression estimator for censored data. Bernoulli. 2013;19(1):344-61.

49. Peng L, Huang Y. Survival analysis with quantile regression models. Journal of the American Statistical Association. 2008 Jun 1;103(482):637-49.

50. Jin Z, Ying Z, Wei LJ. A simple resampling method by perturbing the minimand. Biometrika. 2001 Jun 1;88(2):381-90.

51. Wu QL, J.; Zhu, S.; Wu, J.; Chen, C.; Liu, Q.; Wei, W.; Zhang, Y.; Sun, S.: Breast cancer subtypes predict the preferential site of distant metastases: a SEER based study. Oncotarget 8:27990-27996, 2017

52. Song N, Choi JY, Sung H, Chung S, Song M, Park SK, Han W, Lee JW, Kim MK, Yoo KY, Ahn SH. Heterogeneity of epidemiological factors by breast tumor subtypes in Korean women: a case–case study. International journal of cancer. 2014 Aug 1;135(3):669-81.

53. Dai X, Li Y, Bai Z, Tang XQ. Molecular portraits revealing the heterogeneity of breast tumor subtypes defined using immunohistochemistry markers. Scientific reports. 2015 Sep 25;5:14499.

54. Cade BS, Noon BR. A gentle introduction to quantile regression for ecologists. Frontiers in Ecology and the Environment. 2003 Oct 1;1(8):412-20.