Are relationships between patient demographics and cancer stage different based on Gastrointestinal cancer site locations?

Ari Li[1], Andrew G. Chapple PhD[2,3].

Yale University[1], LSU School of Medicine[2],Stanley S Scott Cancer Center[3]

Background and Data

Many papers have explored associations between patient level factors(i.e. age) and cancer stage when diagnosed. We were interested in testing whether associations differed by cancer sites in Gastrointestinal (GI) cancers. Data collected through the Louisiana Tumor Registry was used for this research project. Patients who were diagnosed with a GI cancer between 2000 and 2020 without prior cancer diagnoses were included.

GI Cancer Site	N	Avg Stage	% 3-4 Stage
Ascending Colon	1206	2.4	39.3
Cecum	1230	2.5	44.6
Esophagus	1246	2.8	62.8
Liver	2062	2.4	45.2
Pancreas	4314	3.2	69.1
Rectum	2204	2.3	41.8
Sigmoid Colon	1684	2.4	42.9
Stomach	1623	2.7	56.9

Table 1: GI cancer site staging information. Sample sizes, average stage, and the % of Cancers with a stage of 3 or 4 are listed.

Demographics of Interest

- Age (mean 66.3)
- Gender (57% Male)
- Race (32% AA)
- Year of Dx (mean 2011)
- Advanced Cancer (55%)
- Smokers (56%)
- High Poverty (41%)
- Rural (23%)
- Private Insurance (28%)
- Ethnicity (2% Hispanic)

Methods

Our goal was to determine whether the relationship between the covariates age, diagnosis year, sex, insurance type, location (urban vs rural), poverty, race, ethnicity, and smoking (denoted X_{1i}, $X_{2i}…X_{9i}$) and stage (denoted Y_i) differed by cancer site using multivariable linear regression. Here, we used the linear regression function (*lm*) in R to model the predicted stage of cancer (Y_i) at diagnosis. The covariates of the model were fit into the regression as a 1 or 0 if the covariate was a categorical variable (i.e. $X_{9i} = 1$ if the patient was a smoker). Dummy variables to indicate the cancer site (D_{2i}, $D_{3i},…D_{8i}$) were also generated to fit the regression model (i.e. $D_{3i} = 1$ if the patients had esophageal cancer). Here the reference category was ascending colon cancers. The assumed regression model without an interaction was modeled as:

$$Y_i = B_0 + \sum_{j=1}^{9} B_j X_{ji} + \sum_{z=2}^{8} \theta_z D_{zi} + \varepsilon_i \quad (1)$$

Here we assume ε_i are of independently and normally distributed. An interaction model was also created for each covariate to test for interactions between cancer site and covariate *m,* which is formally modeled as:

$$Y_i = B_0 + \sum_{j=1}^{9} B_j X_{ji} + \sum_{z=2}^{8} \theta_z D_{zi} + \sum_{z=2}^{8} X_m D_z \delta_z + \varepsilon_i \quad (2)$$

An ANOVA (i.e. *anova* function in R) test was run to determine if the interaction model (2) was significantly different from the base (1) model. P-values <.05 indicated a significant interaction in this nested model test.

Of the significant covariates that had an interaction with cancer site, further analysis was conducted on the site-specific comparisons between the sites and predicted stage. Using the estimated variance covariance matrices (i.e. *vcov* in R), we obtained Wald-based p-values testing whether $H_0: \delta_j = \delta_k$, i.e. that two site-specific interaction effects were equal, via the test statistic.

$$t = \frac{(\hat{\delta}_j - \hat{\delta}_k)^2}{V} \quad \text{where} \quad V = \widehat{\mathrm{Var}}(\hat{\delta}_j) + \widehat{\mathrm{Var}}(\hat{\delta}_k) - 2 * \widehat{\mathrm{Cov}}(\hat{\delta}_j, \hat{\delta}_k).$$

Here t follows a chi-squared distribution. We plotted the significance of the pairwise site-specific interactions (p <0.05) in heatmaps. For each pair of sites, red (blue) indicated a columns site had a more negative (positive) relationship.

Coding

To accomplish these goals, we wrote our own code in R statistical software. Since we tested interactions with 9 different covariate-stage relationships and checked 28 pairwise site effects, we employed these computational devices:

- For Loops
- Logical Statements
- Model & Matrix Indexing

Site-Specific Associations Between Covariance and Cancer Stage

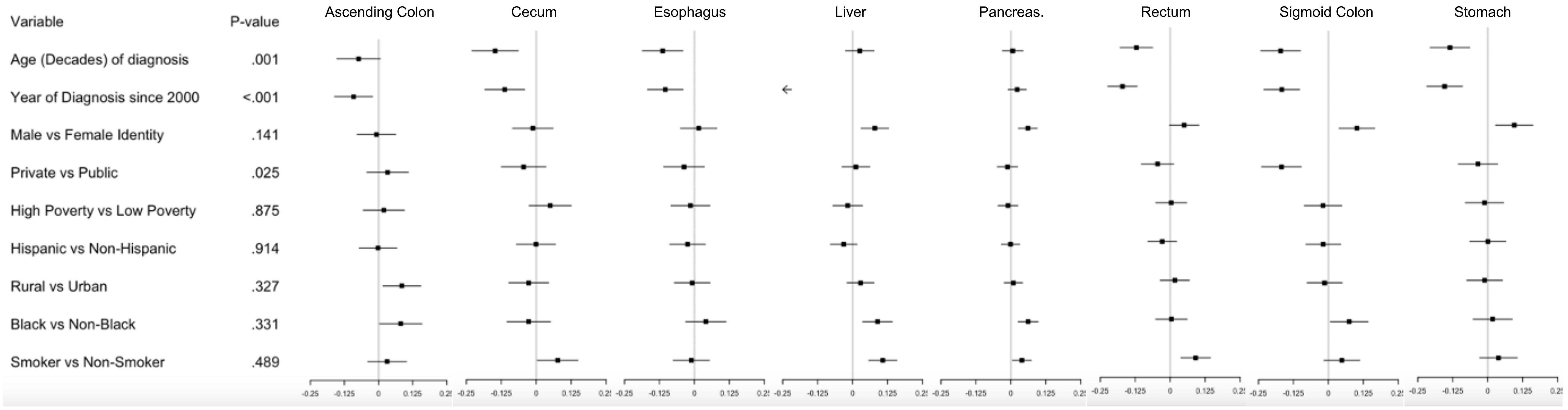

Figure 1. Forest plots of the estimated relationships between each demographic factor and cancer stage within each GI cancer type. For each demographic factor (i.e. private vs public insurance), we display the estimated change in stage along with a 95% confidence interval for this change. Confidence intervals that overlap with 0 indicate a non-significant impact on staging, while confidence intervals greater (less) than 0 indicate a significant increase (decrease) in cancer stage based on that factor. The P-values shown test whether there is a significant difference in these relationships by cancer site.

Exploring the Interactions

Age-Stage Relationship

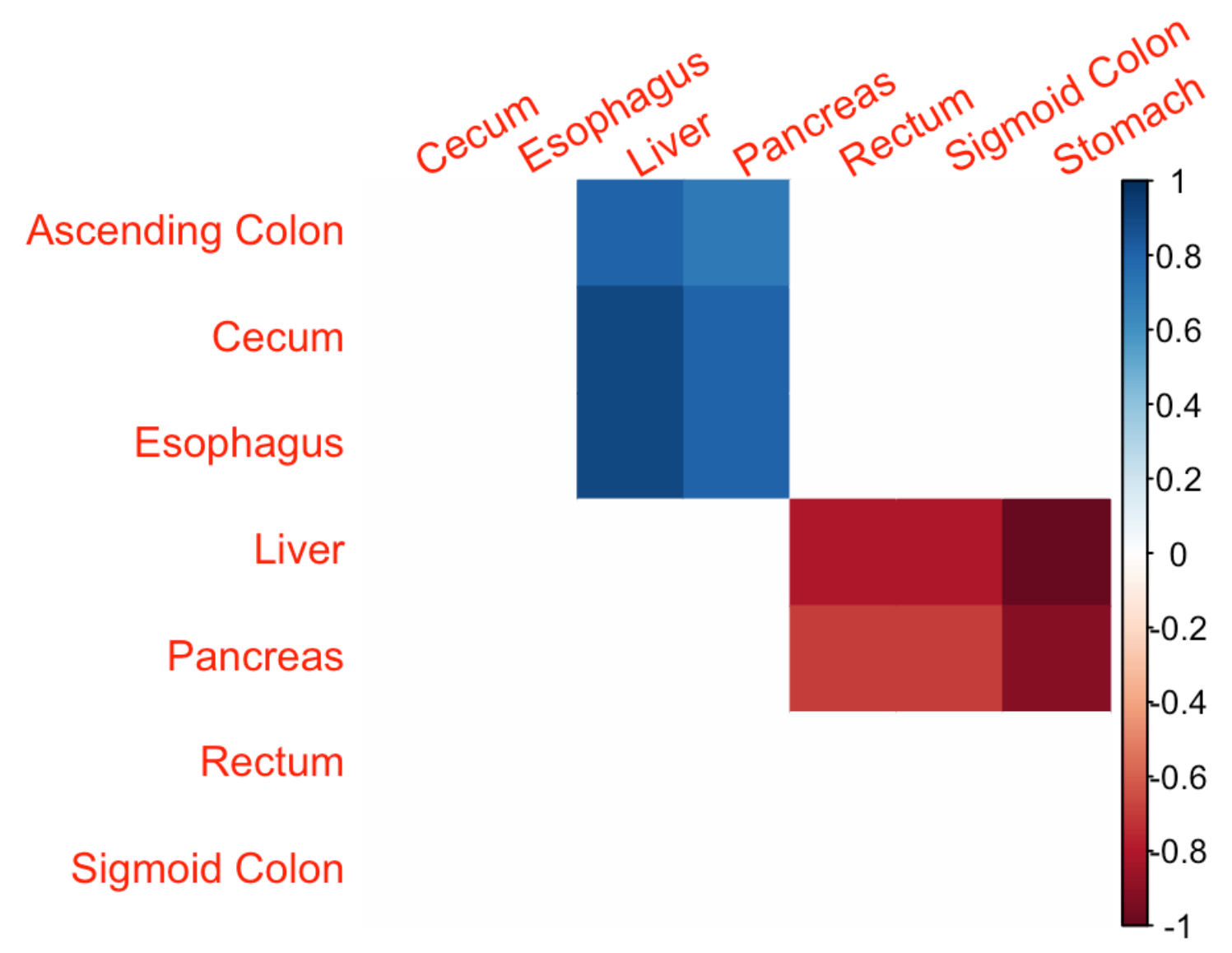

Figure 2. Heatmap showing significant differences in the site-based age-stage relationship. Blue (red) indicates a column site has a more positive (negative) age-stage association than the row site.

Meaning: The heatmap indicates that liver and pancreas cancer have a significantly more positive increase in the age-stage relationship compared to other cancers.

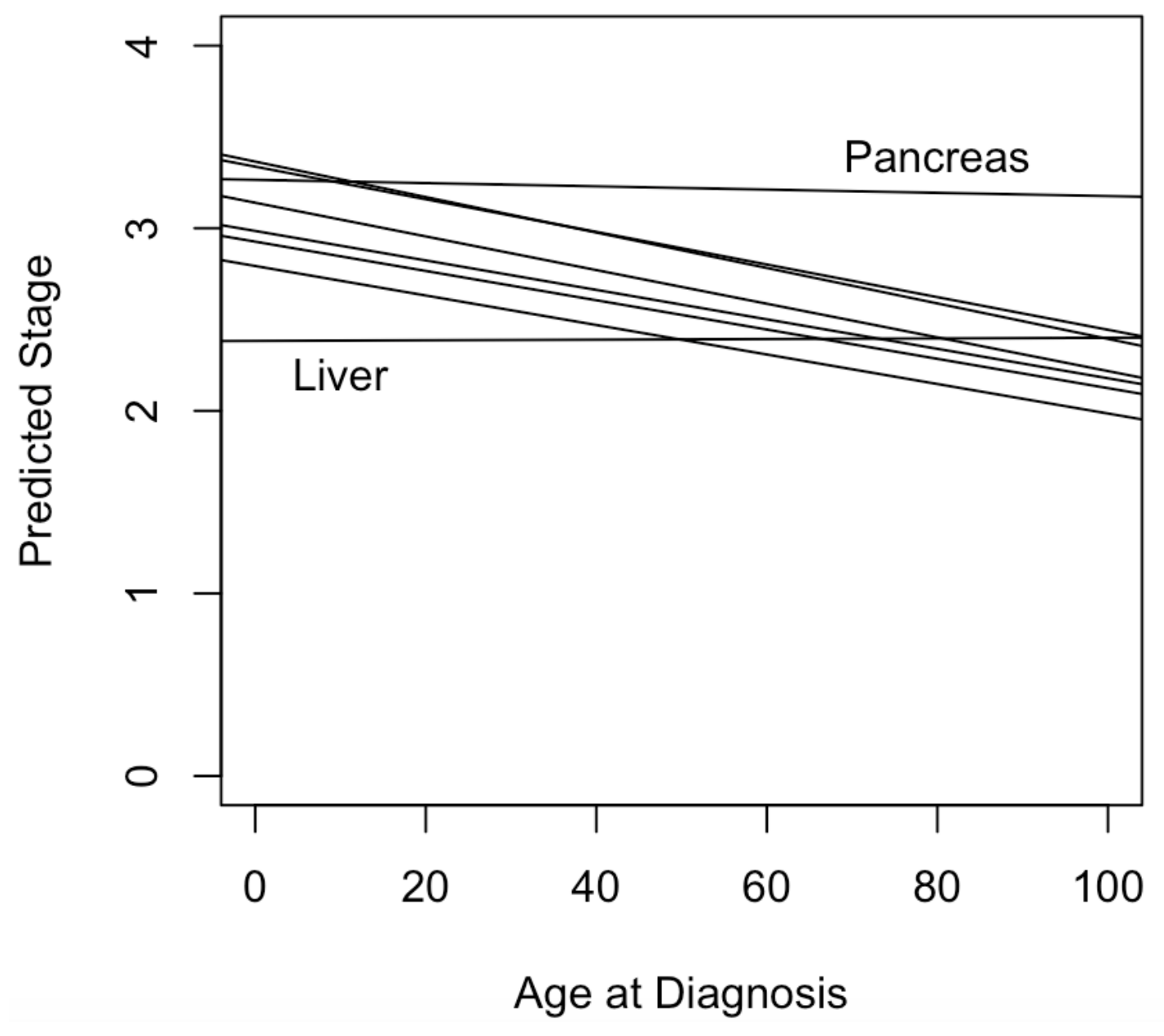

Figure 3. The predicted stage of cancer based on age and average covariate values for year, insurance, etc.

Meaning: Pancreatic and Liver Cancers are shown to be much less influenced by age than other GI cancers.

Year-Stage Relationship

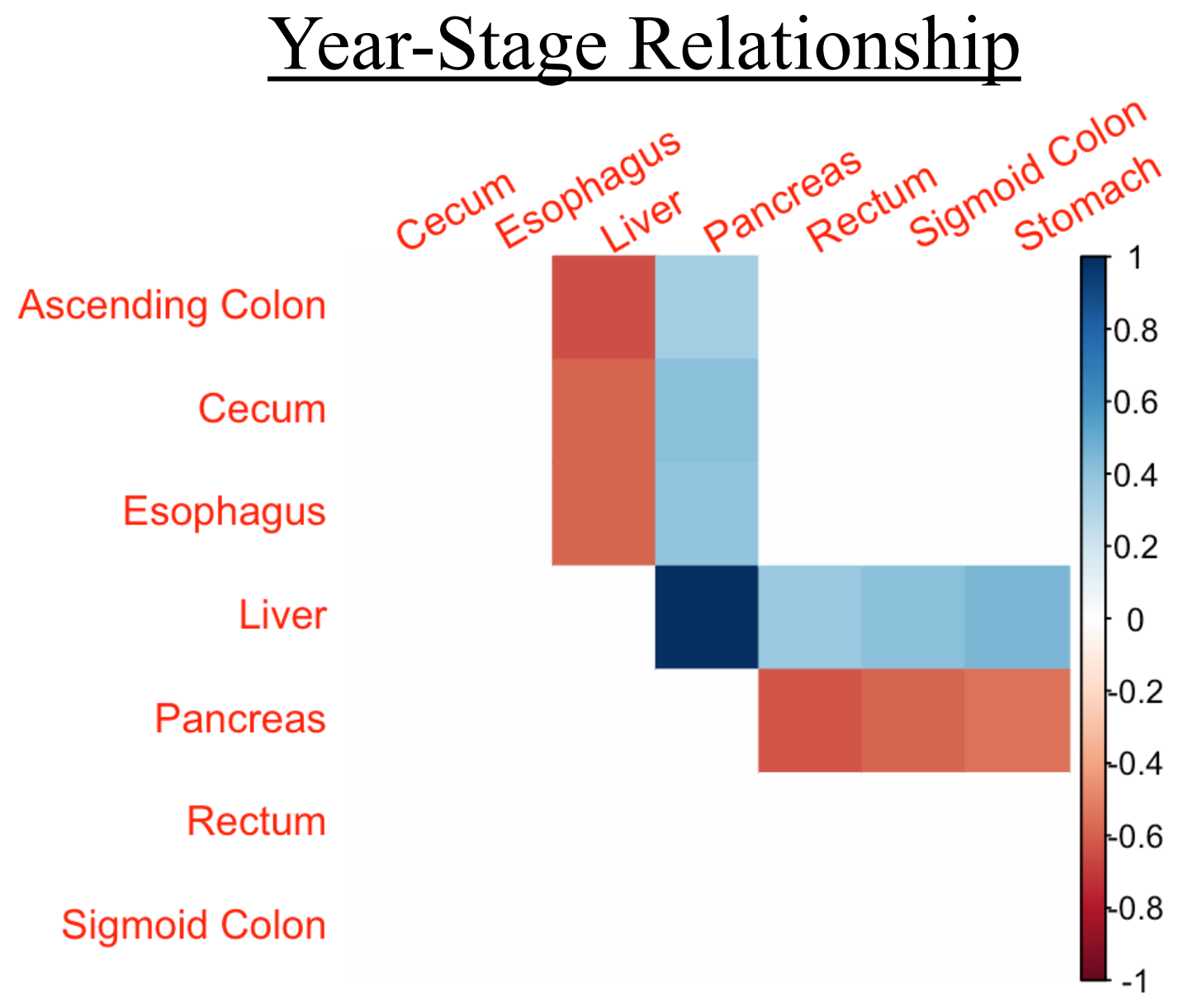

Figure 4. Heatmap showing significant differences in the site-based diagnosis year-stage relationship. Blue (red) indicates a column site has a more positive (negative) age-stage association than the row site.

Meaning: Liver cancer has a significantly more negative age-stage relationship compared to other cancers while the pancreas is significantly more positive.

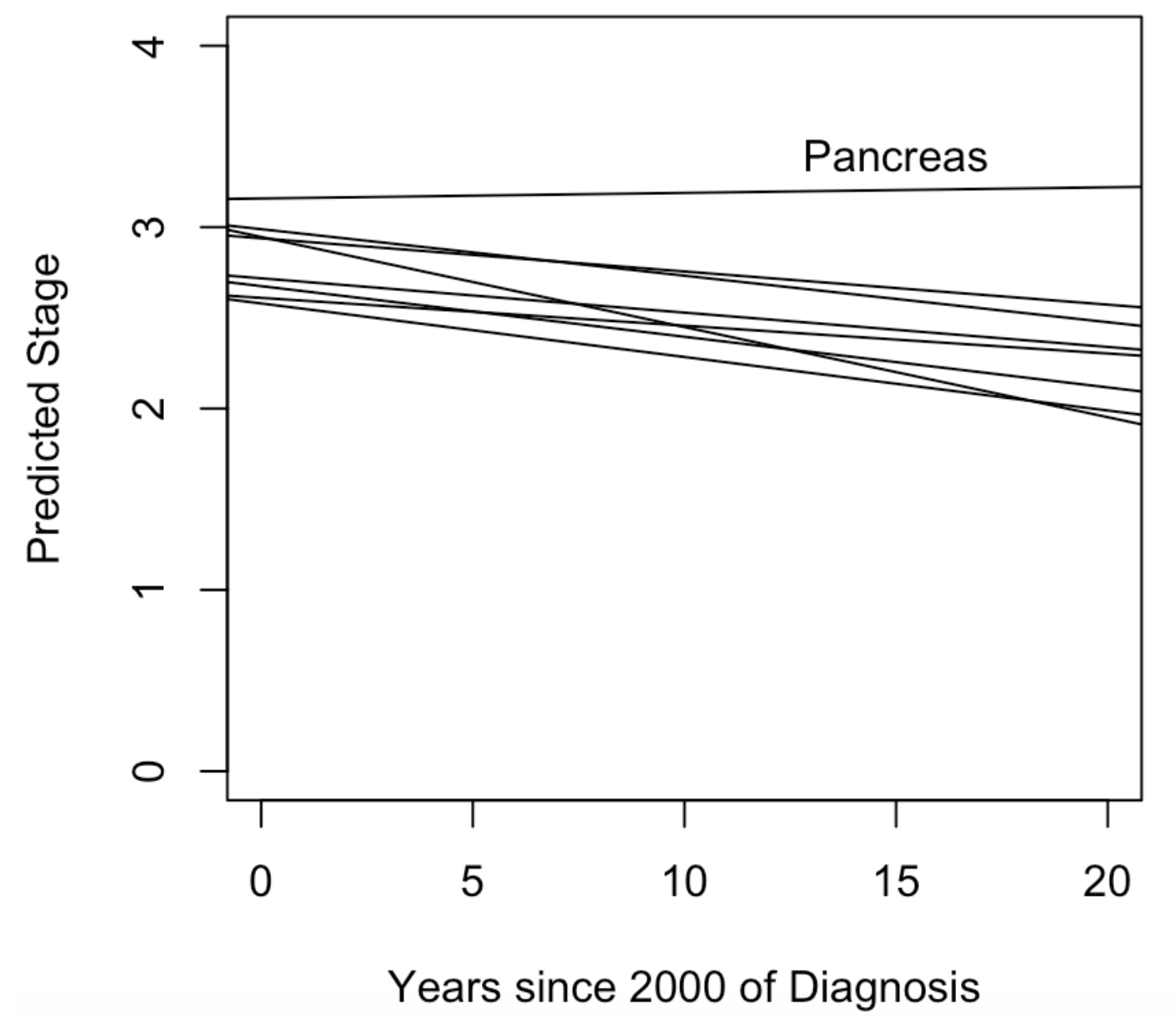

Figure 5. The predicted stage of cancer based on diagnosis year and average covariate values for age, etc.

Meaning: The predicted stage of Pancreatic Cancer is shown to increase in later diagnoses, signaling a decrease in screening efficacy over the years.

Insurance-Stage Relationship

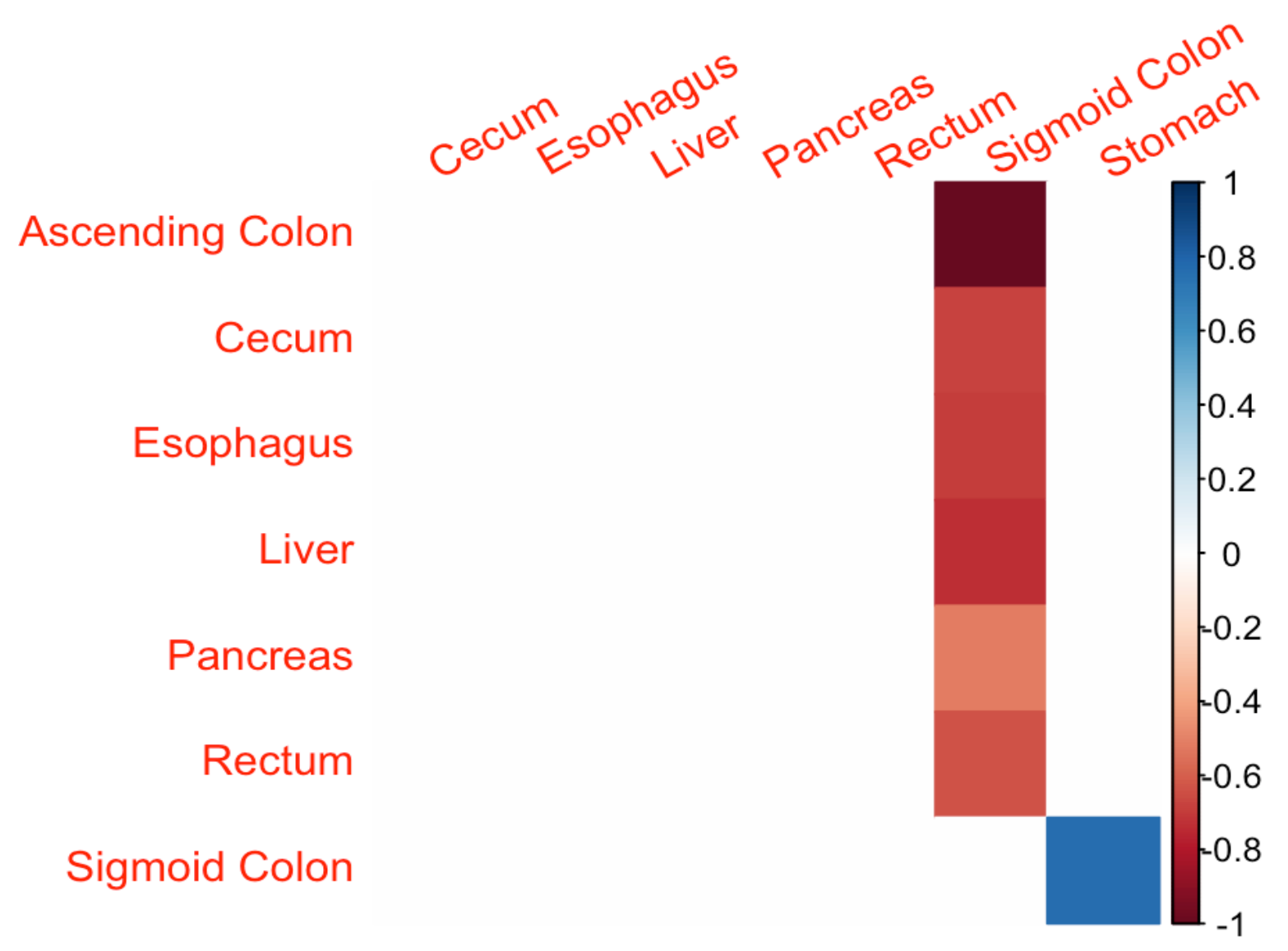

Figure 6. Heatmap showing significant differences in the site-based Insurance-stage relationship. Blue (red) indicates a column site has a more positive (negative) age-stage association than the row site.

Meaning: A strong negative association is depicted between private insurance and Sigmoid Colon cancers when predicting cancer stage compared to other sites.

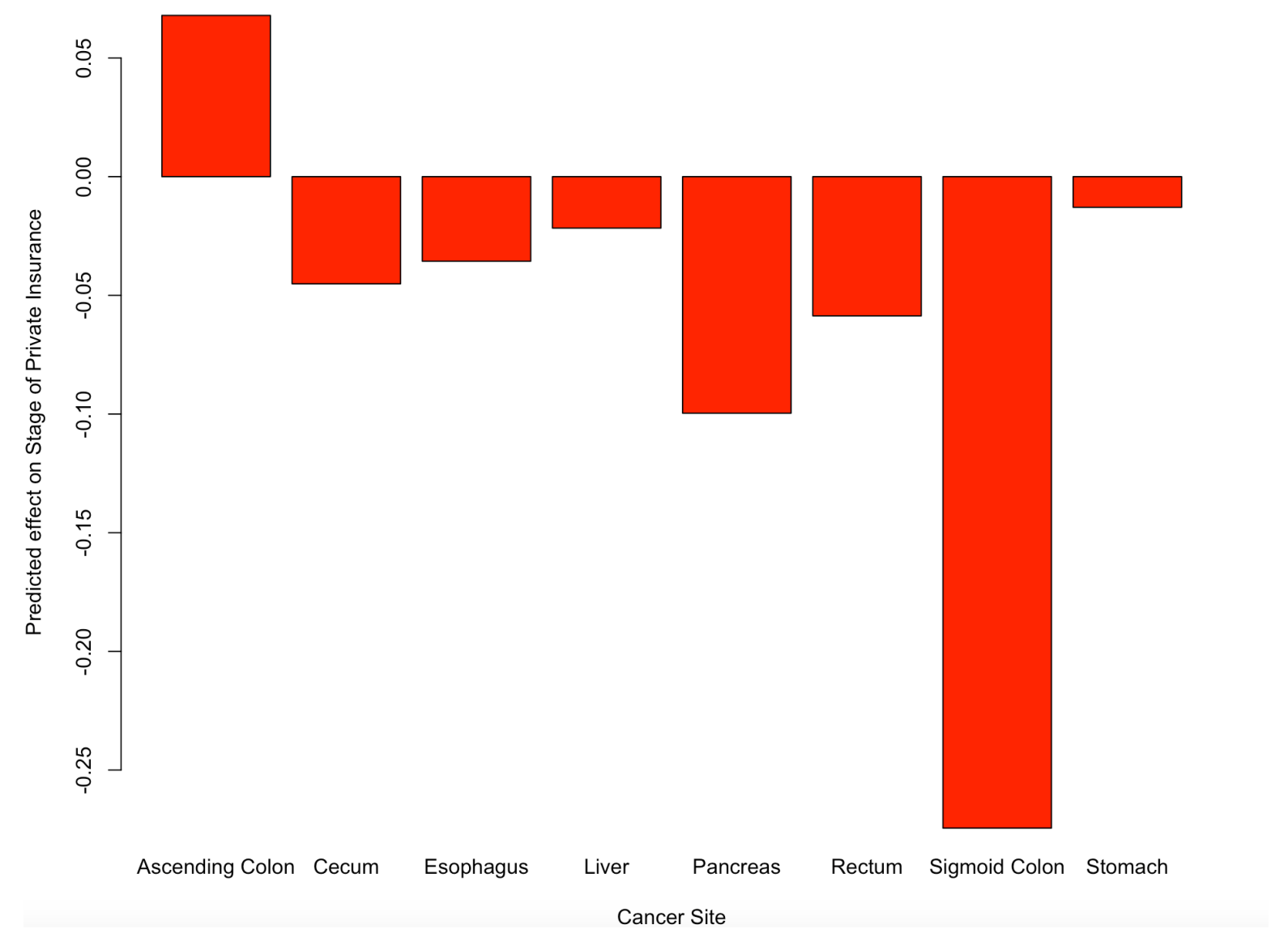

Figure 7. The predicted effect on cancer stage using private vs public insurance and average covariate values.

Meaning: The bar chart shows a drastically decreased predicted stage of Sigmoid Colon for those with private insurance compared to those with public insurance.

This research project was supported through the LSU Health Sciences Center, School of Medicine.