We first used a random forest classification model to predict whether or not a county was above or below the nationwide mean in cancer cost burden (that is, the cancer cost to income ratio).

We used the publicly available Health Rankings dataset from 2018, which included a total of 292 variables (which I will refer to as “features”). We trimmed this dataset down to 152 features (Model 1) that were relatively specific and  “actionable” (or modifiable). For instance, we removed features describing the demographics of each county and, included sub-measures in favor of their less specific summary measures. We also removed features directly related to average incomes, for example poverty levels. In Model 2, we removed features which were correlated at an R value of .9 or greater, keeping the more important feature in each case. For our final model, Model 3, we kept only the top third of features from Model 2 based on importance. You can see that at each step the overall accuracy on the test set improved (from about 80 to 86%), while accuracy on the training set stayed consistent (around 89%). The similar accuracy between the training and testing set in the final model is a good sign and indicates a low degree of overfitting.

On the right, we plot a subset of the feature importances. We see some features that may be expected to contribute to cancer cost burden, such as inactivity, obesity, and smoking, but other interesting results as well, including a high importance of mammography screenings and preventable hospital stays.

Technical points in case asked:

[ Model parameters: forest3 = RandomForestClassifier(n\_estimators = 1000, max\_depth = 10, min\_samples\_leaf = 10) ]

[ We used a train/test set split of 0.9/0.1, but a similar trend in performance/improvement was observed with splits of 0.8/0.2 and 0.75/0.25]

**NEXT SLIDE**

To understand more specifically how well the classifier performed for Tennessee counties, we plotted here the true classifications on the top, and predicted classifications on the bottom. You can appreciate qualitatively how well the model is performing, and ultimately shows an accuracy in Tennessee of 89.5%.

We noted an interesting trend where cancer costs burdens were relatively lower surrounding major metropolitan areas including Memphis, Nashville, and Knoxville.

[Technical note: Not all TN counties are plotted, because some were excluded from modelling due to having a null value in at least one retained feature]

**NEXT SLIDE**

Despite the success of this classification model, we wanted to further validate our findings with another approach, and importantly, delineate the directionality of the relationships between these features and cancer cost burden.

**NEXT SLIDE**

For this we adopted a linear regression approach, using the same features from the final classification model. Importantly, here the predicted variable is the actual cost/income ratio, instead of the binary above/below the national mean variable. This model resulted in a high R-squared and relatively low mean squared error. On the right, we plot the features ranked by their model coefficients. As expected many of the top features are overlapping with those from the classification model, indicating robustness to the type of modelling. But we also get additional information about the directionality of each feature’s effect. For instance, a higher cancer cost burden related to *higher* level of reported physical distress and *lower* levels of mammography screening.

[Technical note: Since this is a simple linear regression model, we aren’t concerned about overfitting. So, the model was fit using all US counties, as opposed to splitting into training and testing sets.]

**NEXT SLIDE**

Again, we plot the actual and predicted cancer cost burdens across Tennessee counties. With few exceptions, the regression model performed well and showed similar performance metrics when looking at only Tennessee counties compared to the full model. This indicates that the trends we observe across all US counties are relatively consistent for Tennessee, and gives us more confidence that our analysis is relevant for TN Med Helper.

I’ll now pass it on to Verra to tell you more about these important factors.

[Technical note: Not all TN counties are plotted, because some were excluded from modelling due to having a null value in at least one retained feature]