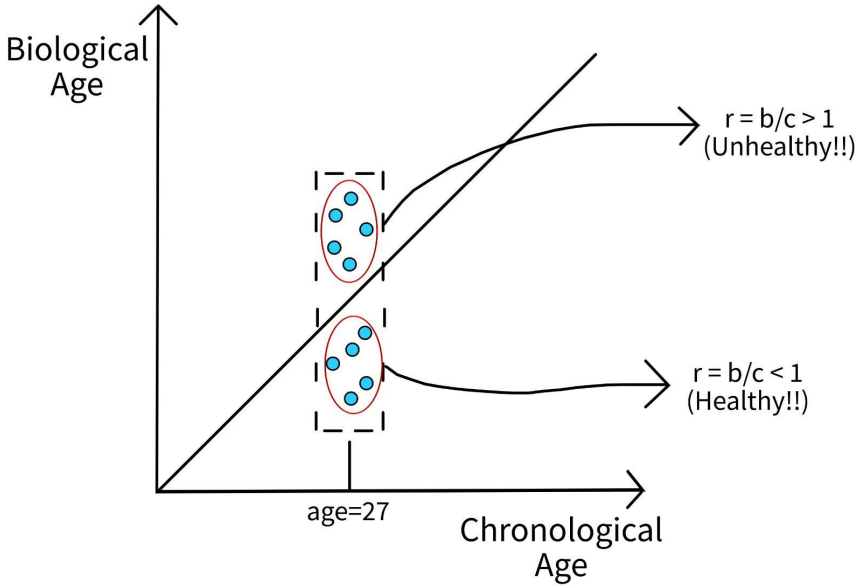
Modeling biological age speedup factor ($bio^{age}/chron^{age}$) as random variable to predict death

Mohil Patel, Sruthi Ganesh, Lipika Garg

Intuition for Modeling



- ➔ Define, $r = bio^{age}/chron^{age}$
- ➔ For a given age, we have healthy and unhealthy people
  - ◆ $r_{age=27} > 1$, implies unhealthy
  - ◆ $r_{age=27} < 1$, implies healthy
- ➔ Intuitively we can expect,
  - ◆ $E[r_{age=27}] = 1$, for that age
- ➔ That is r can be seen a *random variable*, with mean 1
- ➔ We can easily extend the argument to include all ages, so
  - ◆ $E[r] = 1$

Mathematical Model and Assumption-1

Mathematical Model:

- ➔ Define,
  - ◆ $r = bio^{age}/chron^{age}$
- ➔ Model r as a gaussian random variable:
  - ◆ $r = N(1,\sigma^2)$

Assumption-1:

- ➔ Only **1 CT data** available **each patient**, so we can calculate only:
  - ◆ r^{at_CT}
- ➔ To calculate death age, we need:
  - ◆ r^{at_death}
- ➔ Assumption, the value of r stays constant, i.e.:
  - ◆ $r^{at\_CT} = r^{at\_death}$

Note: This assumption can be relaxed if we have more CT data for each patient. In that case we can model: r as a $random\ process$. Or treat r as $time\ series$ & predict r^{at_death}.

Assumption-2

- ➔ To calculate $chron^{age\_at\_death}$ using r, we need $bio^{age\_at\_death}$
- ➔ Assumption:
  - ◆ $bio^{age\_at\_death} = constant \ \& \ same \ for \ everyone$
- ➔ How to calculate $bio^{age\_at\_death}$?
  - ◆ 549 points in dataset with $chron^{age\_at\_death}$ available
  - ◆ Using, $E[r^{at\_death}] = 1$ [From the Model]
  - ◆ $\Sigma(bio^{age\_at\_death} / chron^{age\_at\_death}) / N = 1$ [$bio^{age\_at\_death}$ is constant]
  - ◆ $bio^{age\_at\_death} = harmonic\_mean(chron^{age\_at\_death})$
- ➔ Calculating from the dataset we get:
  - ◆ Fixed, $bio^{age\_at\_death} = 69.35$

Loss Function

- ➔ How to identify whether our predictions of $bio^{age\_at\_death}$ are good or bad?
- ➔ Loss Function (Attempt-1):
  - ◆ $L = (1 - E[r])^2$
  - ◆ $L = (1 - (\Sigma_{all\_datapoint}(r)/N))^2$

Issues:

- ➔ Some ages can have $r > 1$, and others may have $r < 1$
- ➔ But averaging across all ages cancel things out
- ➔ Optimizing across all ages

- ➔ Loss Function (Better Alternative):
  - ◆ $L = \Sigma_{age}[(1 - E[r_{age}])^2]$

Advantages:

- ➔ Ensuring **each age** will have $E[r_{age}] = 1$
- ➔ *No across age averaging*

Approaches

➔ $r^{at\_CT} = r^{at\_death}$ [Assumption-1]

➔ $(bio^{age\_at\_CT} / chron^{age\_at\_CT}) = (bio^{age\_at\_death} / chron^{age\_at\_death})$ [eq-1]

Unknown!!

- ➔ 2 unknowns in the equation
- ➔ We will define one using an ML model
- ➔ Other can be found using the equation

Unknown!!

Approach-1:

- ➔ Use ML Model to predict $chron^{age\_at\_death}$
- ➔ Calculate $bio^{age\_at\_CT}$ using the **eq-1**
- ➔ Validate $bio^{age\_at\_CT}$ values using loss function defined earlier (reflects how good the model is)

Issues:

- ➔ Only 549 points with data labels
- ➔ Due to **data limitation** trained **ML models** will **not have good accuracy**

Approaches (Continued)

Approach-2 (better approach):

- ➔ Define $bio^{age\_at\_CT}$ using ML model (we have tried different models and will discuss them in next slides)
- ➔ Validate the goodness of the model using the loss function we defined earlier
- ➔ Get $chron^{age\_at\_death}$ using **eq-1**, i.e.:
  - ◆ $chron^{age\_at\_death} = (bio^{age\_at\_death} \times chron^{age\_at\_CT}) / bio^{age\_at\_CT}$

Advantages of using approach-2:

- ➔ Have 9223 data points to work with
- ➔ Can define $bio^{age\_at\_CT}$ in many different ways (Be Creative!!)

Note: Due to time constraints, we will only discuss **approach-2 in this presentation**. For **approach-1 results** please refer to the **report**

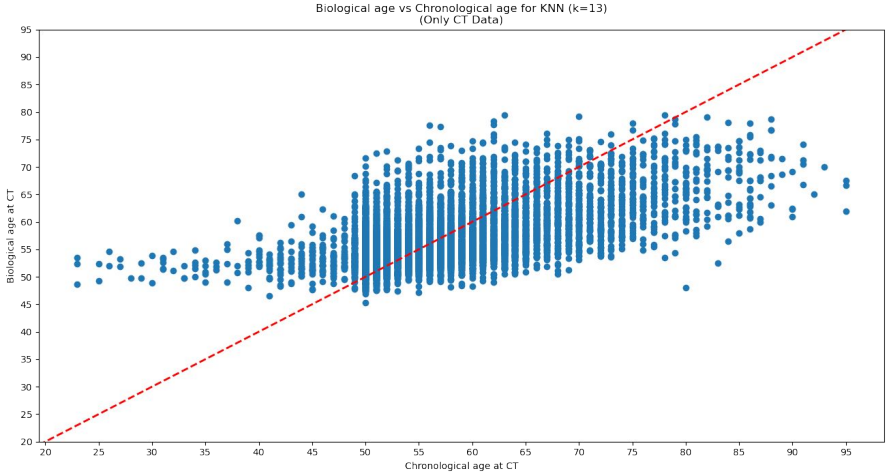
Model-1: KNN

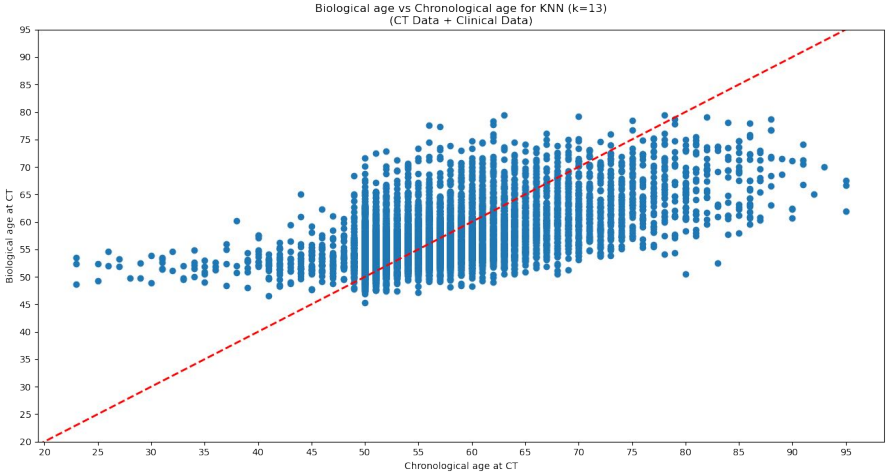
Key Idea: Nearest Neighbors using CT (& Clinical Data) should have similar $bio^{age\_at\_CT}$

➔ $E[r] = 1$ [From Modeling]

➔ $E[bio^{age\_at\_CT}/chron^{age\_at\_CT}] = 1$ [Using $bio^{age\_at\_CT}$ as constant from key idea]

➔ $bio^{age\_at\_CT} = E[chron^{age\_at\_CT}]$ → Take average age of k-nearest neighbors, treat as $bio^{age\_at\_CT}$



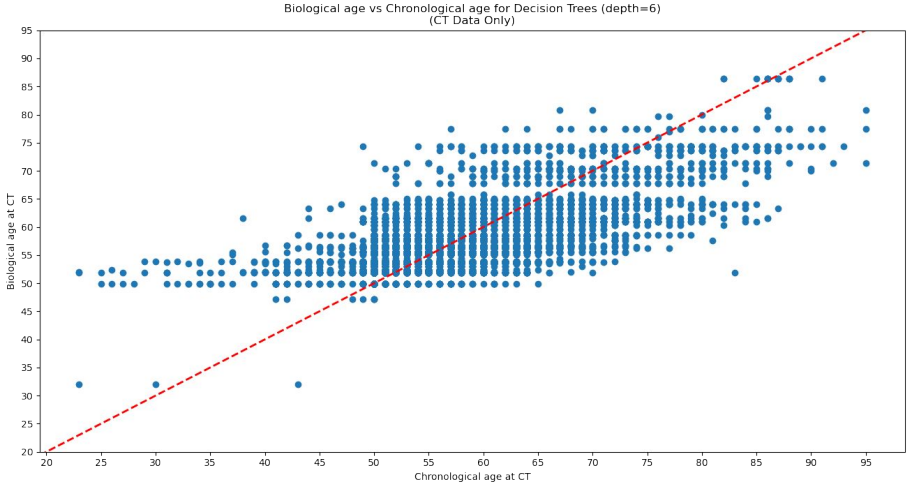


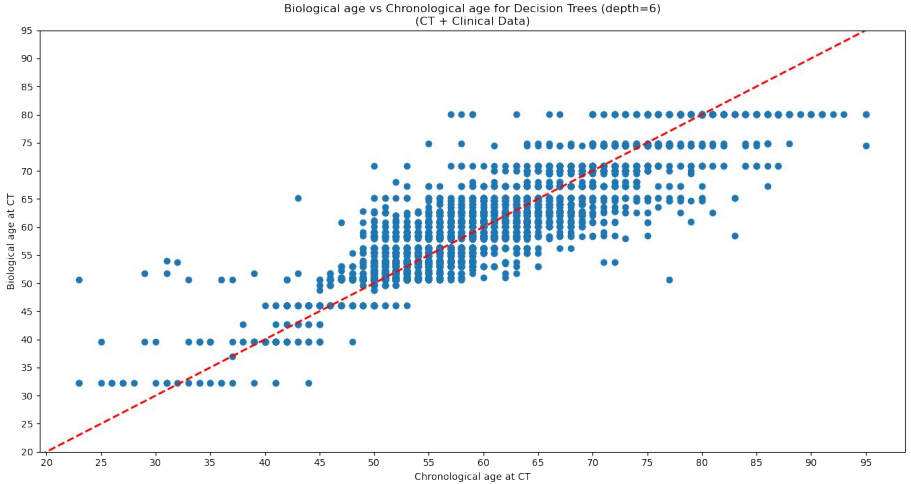
Loss Score (CT Only) = 10.16 → *Improvement - 0.1%* → *Loss Score (CT + Clinical) = 10.15*

Model-2: Regression Decision Tree

Key Idea: Similar as KNN. $bio^{age\_at\_CT}$ is same if similar CT values.

- ➔ Similar CT values → Same Leaf Node in Decision Tree
- ➔ That implies, Patients with same $bio^{age\_at\_CT}$ → Same Leaf Node
- ➔ Thus we can treat, $bio^{age\_at\_CT} = Result\_of\_Regression\_Decision\_Tree(CT\_Values)$





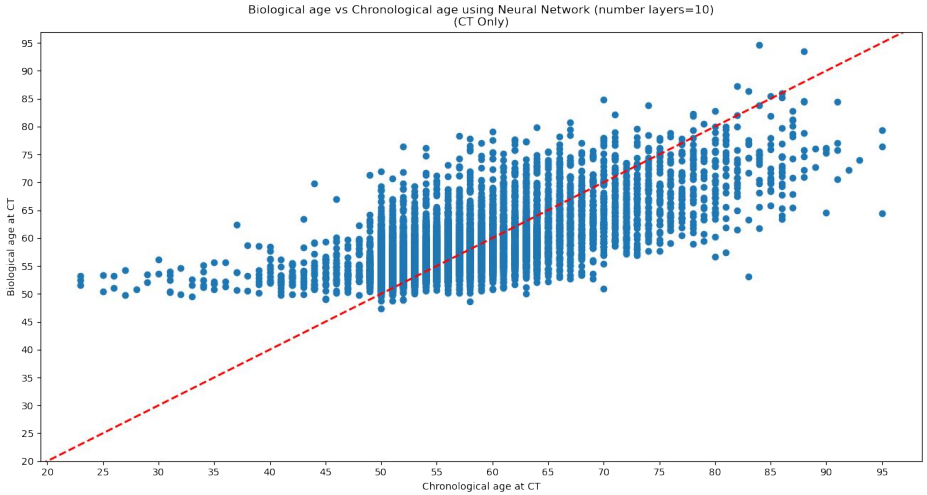
Loss Score (CT Only) = 8.47 → *Improvement - 80.8%* → *Loss Score (CT + Clinical) = 1.62*

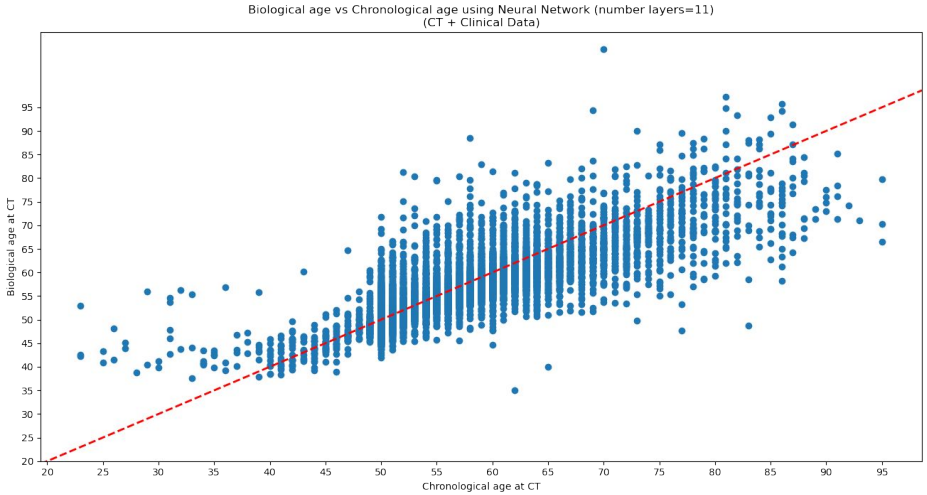
Model-3: Neural Network

Key Idea: Similar as last 2 models:

➔ For a given CT Values → find characteristics age

➔ We treat that age as $bio^{age\_at\_CT} = Result\_of\_Neural\_Network(CT\_Values)$





Loss Score (CT Only) = 10.29 → *Improvement - 55.4%* → *Loss Score (CT + Clinical) = 4.59*

Alternative idea to define $bio^{age\_at\_CT}$

Till now, the models we discussed are based on following fundamental idea:

➔ For given CT values → find characteristics age → that characteristic age is $bio^{age\_at\_CT}$

Why don't we reverse the approach?

➔ For a given age → find characteristic CT values → Make a lookup table for all the ages

For any new CT value:

➔ Use the lookup table and find the nearest point

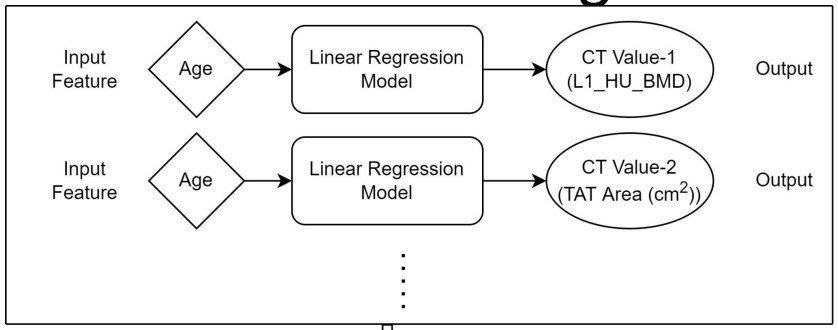
➔ The age corresponding to point is defined as $bio^{age\_at\_CT}$

The fundamental idea that we want to cover is:

➔ For every age, there is a characteristic CT value which defines that age

Age	**Characteristic CT Value - 1**	**Characteristic CT Value - 2**	**…….**	**Characteristic CT Value - k**
1	$v_1 1$	$v_1 2$	…….	$v_1 k$
2	$v_2 1$	$v_2 2$	…….	$v_2 k$
…….				
100	$v_{100} 1$	$v_{100} 2$	…….	$v_{100} k$

Model-4: Linear Regression

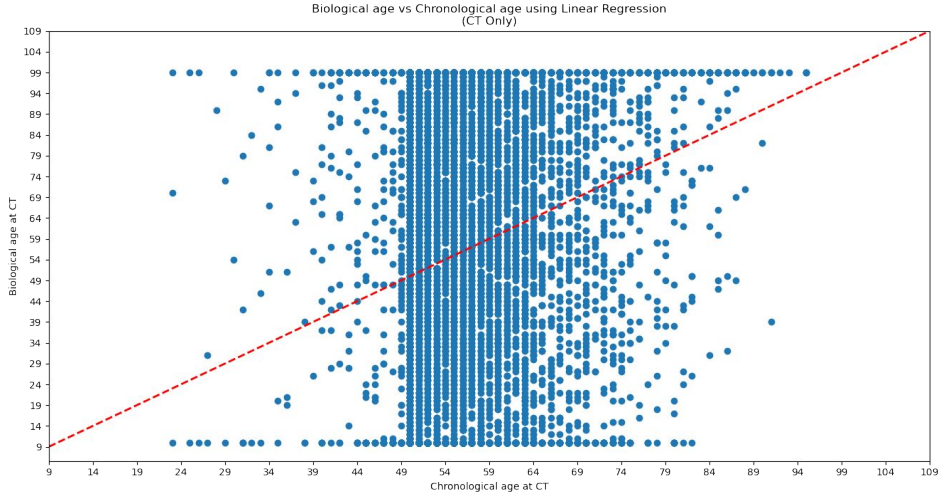


Lookup Table

Age	CT-Val1	CT-Val2	CT-Val3	.....
n	Vn1	Vn2	Vn3	.....
...	...	...	...	.....

Lookup Table:

- ➔ Has characteristic CT values for all the ages
- ➔ New CT Value → Find nearest point in lookup table → Define that age as $bio^{age\_at\_CT}$



Loss Score (CT Only) = 20.36

Loss score of the above model is not too bad. But clearly the plot above shows that the model is bad. This leads to identifying some issues:

- ➔ Picking single value from the lookup table might not be the best way (need some different approach)
- ➔ Loss score does not capture variance as optimization goal

Conclusion

Takeaways:

- ➔ defined a mathematical modeling for $r = bio^{age}/chron^{age}$
- ➔ defined a loss function to access goodness of fit: $L = \Sigma_{age}[\ (1 - E[r_{age}])^2\ ]$
- ➔ implemented multiple ML models and accessed their accuracy
- ➔ regression decision trees are showing good results
- ➔ adding clinical data is improving results for all the ML models

Future Works:

- ➔ incorporate variance in the loss function to make optimization goal better (previous slide!!)
- ➔ currently we assume: $r^{at\_CT} = r^{at\_death}$, because only 1 CT value available. Extend this to a better mathematical model if more than 1 CT values available for each patients
  - ◆ r as $random\ process$
  - ◆ treating r as a $time\ series$ prediction
- ➔ make an estimator for variance (in $r = N(1,\sigma^2)$) and try to do confidence interval estimation