Data distribution and error scores on imputation methods
Summary

Findings on our assessed prostate cancer data

• Data distributions before/after imputing missingness are preserved for most of the cases

• Data with similar types (binary, discrete, continuous) behave similarly towards different imputation methods

• The error scores of univariate methods are often comparable with multivariate methods
MISSINGNESS AND RMSE PER FEATURE (COMPLETE SET)

<table>
<thead>
<tr>
<th>Feature</th>
<th>Missingness</th>
<th>RMSE - BEST</th>
<th>RMSE - MICE</th>
<th>RMSE - KNN</th>
</tr>
</thead>
<tbody>
<tr>
<td>BCR PSA</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>BCR status</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Post PSA</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>BCR time</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Tumor margin</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Pre-OP therapy</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Data Distribution
Before/After Imputation
PRE-OP THERAPY – NUMBER OF VALUES PER CATEGORY

ORIGINAL
- 2 missing
- 10 no
- 4 ADT
- 1 Chemo
- 1 TRUP
- 1 Radio bone

IMPUTED
- 10 no
- 4 ADT
- 1 Chemo
- 1 TRUP
- 1 Radio bone
TUMOR MARGIN – NUMBER OF VALUES PER CATEGORY

<table>
<thead>
<tr>
<th>Category</th>
<th>0</th>
<th>1</th>
<th>Missing</th>
</tr>
</thead>
<tbody>
<tr>
<td>ORIGINAL</td>
<td>76</td>
<td>49</td>
<td>7</td>
</tr>
<tr>
<td>BEST/MICE</td>
<td>83</td>
<td>49</td>
<td>0</td>
</tr>
<tr>
<td>KNN</td>
<td>80</td>
<td>52</td>
<td>0</td>
</tr>
</tbody>
</table>
BCR STATUS – NUMBER OF VALUES PER CATEGORY

<table>
<thead>
<tr>
<th>Method</th>
<th>0</th>
<th>1</th>
<th>Missing Values</th>
</tr>
</thead>
<tbody>
<tr>
<td>Original</td>
<td>44</td>
<td>49</td>
<td>48</td>
</tr>
<tr>
<td>Best</td>
<td>92</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td>MICE</td>
<td>78</td>
<td>53</td>
<td></td>
</tr>
<tr>
<td>KNN</td>
<td>76</td>
<td>56</td>
<td></td>
</tr>
</tbody>
</table>

ORIGINAL: 48 missing values, BEST: 92, MICE: 53, KNN: 56
Root Mean Square Error (RMSE) of imputation methods for simulated missingness in the data
RMSE – TUMOR MARGIN (TRAINING SET)
Root Mean Square Error – POST-PSA

Percentage of Missing Values

RMSE
RMSE – BCR STATUS

- MICE
- BEST

PERCENTAGE OF MISSING VALUES

- Training set
- Test set
- Complete set

BEST

RMSE

RMSE – BCR STATUS

PERCENTAGE OF MISSING VALUES

5% 10% 15% 20% 25% 30% 35% 40% 45% 50% 55% 60% 65% 70% 75% 80% 85% 90% 95%
Binary Features
TUMOR MARGIN (TEST SET)

A .. Most frequent, mean, median, linear regression, multivariate, KNN
B .. Binary features: Multivariate, MICE, KNN
C .. Numeric features: Multivariate, MICE, KNN
TUMOR MARGIN (COMPLETE SET)

A .. Most frequent, mean, median, linear regression, multi variate, KNN
B .. Binary features: Multi variate, MICE, KNN
C .. Numeric features: Multi variate, MICE, KNN
BCR STATUS (TRAINING SET)

- A: Most frequent, mean, median, linear regression, multivariate, KNN
- B: Binary features: Multivariate, MICE, KNN
- C: Numeric features: Multivariate, MICE, KNN
BCR STATUS (TEST SET)

A .. Most frequent, mean, median, linear regression, multi variate, KNN
B .. Binary features: Multi variate, MICE, KNN
C .. Numeric features: Multi variate, MICE, KNN
BCR STATUS (COMPLETE SET)

A .. Most frequent, mean, median, linear regression, multi variate, KNN

B .. Binary features: Multi variate, MICE, KNN

C .. Numeric features: Multi variate, MICE, KNN
Continuous Features
POST PSA (TRAINING SET)

A .. Median
B .. Most frequent
C .. Continuous features: multivariate, MICE, KNN
D .. Mean, linear regression, multivariate, KNN
E .. Numeric features: multivariate, MICE, KNN
POST PSA (TEST SET)

A .. Median
B .. Most frequent
C .. Continuous features: multi variate, MICE, KNN
D .. Mean, linear regression, multi variate, KNN
E .. Numeric features: multi variate, MICE, KNN
POST PSA (COMPLETE SET)

A .. Median
B .. Most frequent
C .. Continuous features: multi variate, MICE, KNN
D .. Mean, linear regression, multi variate, KNN
E .. Numeric features: multi variate, MICE, KNN
BCR PSA (TRAINING SET)

A .. Median
B .. Most frequent
C .. Continuous features: multi variate, MICE, KNN
D .. Mean, linear regression, multi variate, KNN
E .. Numeric features: multi variate, MICE, KNN
BCR PSA (TEST SET)

A .. Median
B .. Most frequent
C .. Continuous features: multi variate, MICE, KNN
D .. Mean, linear regression, multi variate, KNN
E .. Numeric features: multi variate, MICE, KNN
BCR PSA (COMPLETE SET)

A .. Median
B .. Most frequent
C .. Continuous features: multi variate, MICE, KNN
D .. Mean, linear regression, multi variate, KNN
E .. Numeric features: multi variate, MICE, KNN
Discrete Features
BCR TIME (TRAINING SET)

A .. Mean, linear regression, multi variate, KNN
B .. Median
C .. Discrete features: multi variate, MICE, KNN
D .. Most frequent
E .. Numeric features: multi variate, MICE, KNN
BCR TIME (COMPLETE SET)

PERCENTAGE OF MISSING VALUES

RMSE

A .. Mean, linear regression, multi variate, KNN

B .. Median

C .. Discrete features: multi variate, MICE, KNN

D .. Most frequent

E .. Numeric features: multi variate, MICE, KNN