FDA U.S. FOOD & DRUG ADMINISTRATION

Sample size determination for electronic phenotyping studies

Satabdi Saha, Sai Dharmarajan, JaeJoon Song

U.S. Food and Drug Administration, Center for Drug Evaluation Research (CDER)

1. Background

Advances in automated document classification has led to identifying massive numbers of clinical concepts from handwritten clinical notes. These high dimensional clinical concepts can serve as highly informative predictors in building classification algorithms for identifying patients with different clinical conditions, commonly referred to as patient phenotyping. However, from a planning perspective, it is first critical to ensure that enough data is available for the phenotyping algorithm to obtain a desired classification performance. This challenge in sample size planning is further exacerbated by the high dimensionality of the covariates and the inherent imbalance of the response class. In this poster we describe a two-step approach for sample size planning. In Step 1, we show how to incorporate feature selection in a linear discriminant analysis using two different approaches. Then, in Step 2, we derive formulas for sample size requirements based on optimizing classification performance metrics sensitive to class imbalance (AUC, MCC). Therefore, our method determines sample size for a linear classifier incorporating feature selection.

2. Methods

Step 1: High Dimensional Feature Selection.

- We consider the two-class classification problem with the high dimensional covariate vector $x \in N(\mu, \Sigma)$ when $x \in C_1$ and $x \in N(-\mu, \Sigma)$ when $x \in C_2$; $y_i = I\ (x \in\ C_1)$.
- Using LDA classify $x \in C_1$ when $\boldsymbol{2x^T\Sigma^{-1}\mu > \kappa \Rightarrow w^T x > \kappa}$ where $\kappa = \log\left(\frac{1-p_1}{p_1}\right)$
- Given the high-dimensionality of the feature space we employ a feature selection procedure to eliminate $(p - m)$ redundant covariates, hence making Σ non-singular.
- Only the remaining m features are included in the linear classifier.
- Dobbin Simon (DS) method[1] employs a two-sample t-test to select features that are significant for a given pilot data
- HCT method employs Higher Criticism Thresholding[2] approach to select m important features out of the p total features

Step 2: Computation of sample size dependent performance metrics.

- On obtaining the m important features, performance accuracy metrics sensitive to imbalanced class datasets .are derived both the DS and HCT approaches
- **DS Method**
  - Define $\theta = (\delta, m, \beta, \lambda, p, \kappa)$ where δ denotes the minimum effect size, m is the total number of important features, p is the total number of features, β is the power of the test, α is the level of the test, λ is the maximum eigen value of the population correlation matrix and
  - $\boldsymbol{AUC(n) = \int_{\kappa=-\infty}^{\kappa=\infty} Q_1(\theta) d(1 - Q_2(\theta))}$ **, where:**

$$TPR = Q_1(\theta) = \Phi\left(\frac{\delta m(1-\beta) - \kappa}{\sigma\sqrt{(\lambda)\sqrt{m(1-\beta) + (p-m)\alpha}}}\right)$$

$$TNR = Q_2(\theta) = \Phi\left(\frac{\delta m(1-\beta) + \kappa}{\sigma\sqrt{(\lambda)}\sqrt{(m(1-\beta) + (p-m)\alpha)}}\right)$$

$$Precision = Q_3(\theta) = \frac{(p_1 \times Q_1(\theta))}{(p_1 \times Q_1(\theta)) + ((1-p_1) \times (1 - Q_2(\theta)))}$$

$$NPV = Q_4(\theta) = \frac{((1-p_1) \times Q_2(\theta))}{((1-p_1) \times Q_2(\theta)) + (p_1 \times (1 - Q_1(\theta)))}$$

$$\boldsymbol{MCC(n) = \sqrt{Q_1(\theta) \times Q_2(\theta) \times Q_3(\theta) \times Q_4(\theta)} - \sqrt{(1 - Q_1(\theta)) \times (1 - Q_2(\theta)) \times (1 - Q_3(\theta)) \times (1 - Q_4(\theta))}}$$

- **HCT Method**
  - Compute Z-scores $Z = \left(\frac{1}{S_\Delta}\right)\sum_{i=1}^{n} X_i Y_i$; $S_\Delta = \sqrt{\left(\frac{s_1^2}{n_1} + \frac{s_2^2}{n_1}\right)}$
  - For an appropriately chosen HCT threshold λ , $w_j = sign(Z_j).I(|Z_j| > \lambda)\ and\ w_j = 0\ otherwise$

$$\boldsymbol{AUC(n) = \int_{\kappa=-\infty}^{\kappa=\infty} \Phi\left(\frac{w^T\mu - \kappa}{w^T\Sigma^{-1}w}\right) d\left(1 - \Phi\left(\frac{w^T\mu - \kappa}{w^T\Sigma^{-1}w}\right)\right)}$$

3. Real Data Analysis to Evaluate the Methods Performance

3.1. Data Description:

Clinical notes were extracted from the MIMIC-III database[3-5] which contains de-identified clinical data of over 53,000 hospital admissions for adult patients to the intensive care units (ICU) at the Beth Israel Deaconess Medical Center from 2001 to 2012. This project uses a dataset of 833 patient discharge summaries restricted to frequently readmitted patients (>3 in a single year), labeled with 15 clinical patient phenotypes believed to be associated with risk of recurrent readmission by domain experts.

3.2. Phenotyping Task 1:

- Focused on building a classifier for identifying patients with *'Depression'*, which had a prevalence of 29%
- Notes were transformed into Unified Medical Language System (UMLS) Concepts using MetaMap Lite. Each note represented as a vector of 10,109 Concepts
- Performance from a LASSO using 80-20 training-test split was compared to estimated performance from our proposed sample size determination methods.
- Performance on Matthew's Correlation Coefficient (MCC) and Area under the ROC curve (AUC) was estimated for varying sample sizes n and number of important features m using methods described.

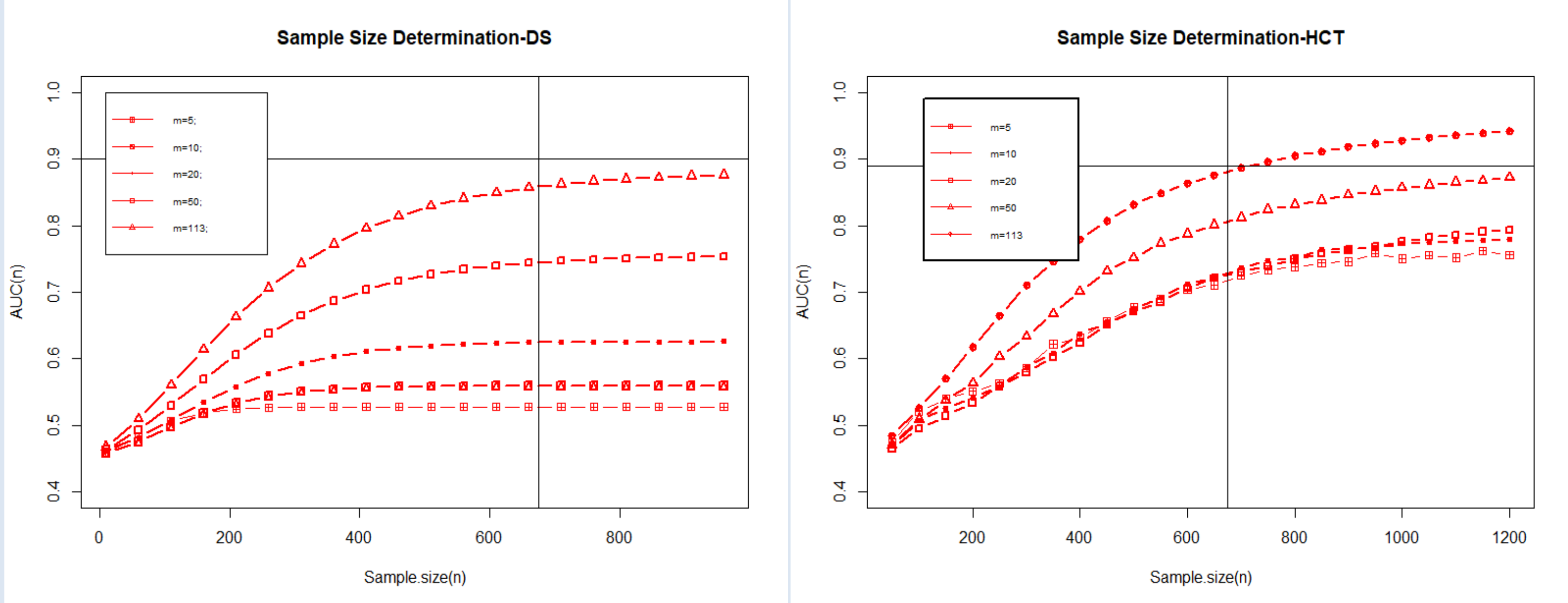


Figure 1: Sample Size Determination using DS and HCT for independent features.

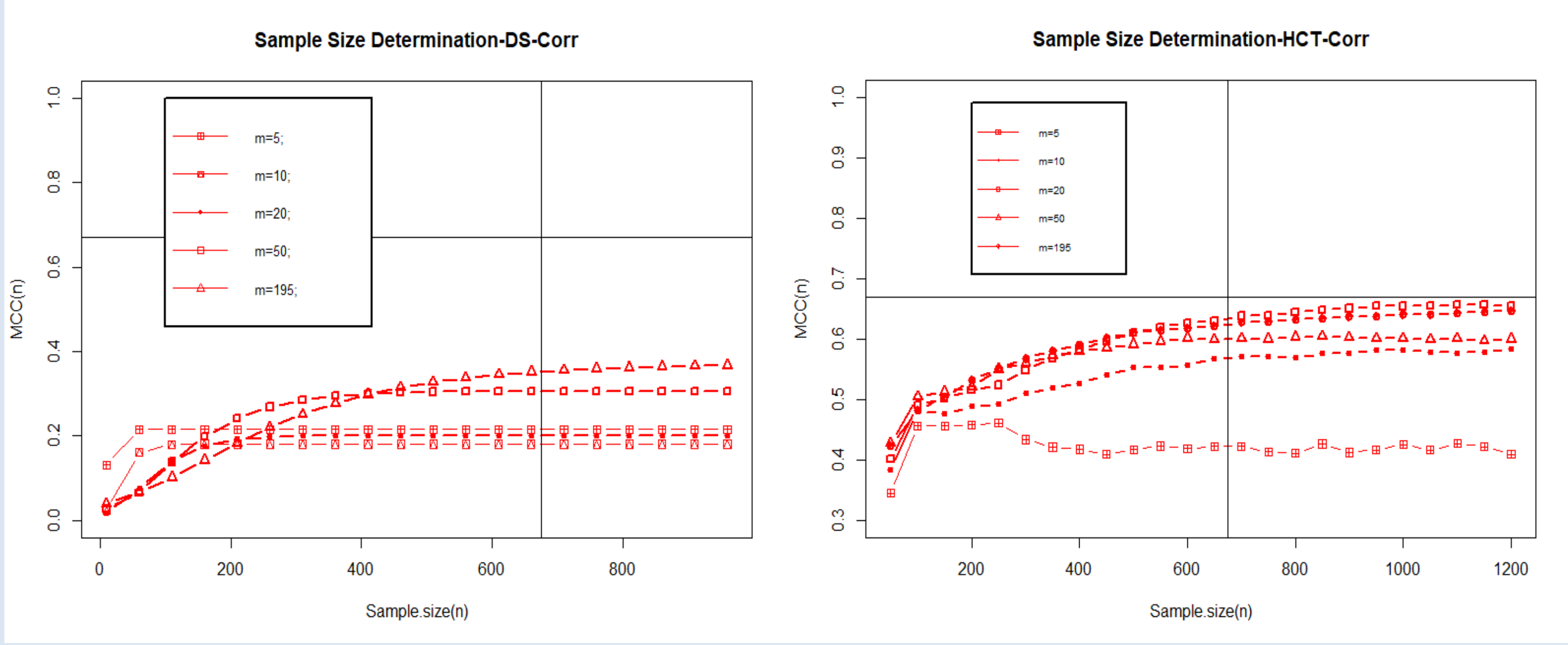


Figure 2: Sample size determination using DS and HCT for correlated features.

4. Conclusions

- DS and HCT methods can provide a reasonable estimate of the sample size required for a linear ML classifier like LASSO.
- Our experience suggests these would be conservative estimates; perhaps even more so for a more non-linear ML algorithm like random forest, ANN
- HCT method most preferable as it does not require features to be independent.

5. References

1. Dobbin, Kevin K., and Richard M. Simon. "Sample size planning for developing classifiers using high-dimensional DNA microarray data." Biostatistics 8, no. 1 (2007): 101-117
2. Donoho, David, and Jiashun Jin. "Higher criticism for large-scale inference, especially for rare and weak effects." Statistical Science 30, no. 1 (2015): 1-25.
3. Moseley, E., Celi, L. A., Wu, J., & Dernoncourt, F. (2020). Phenotype Annotations for Patient Notes in the MIMIC-III Database (version 1.20.03). PhysioNet.
4. Gehrmann S, Dernoncourt F, Li Y, Carlson ET, Wu JT, Welt J, et al. (2018) Comparing deep learning and concept extraction based methods for patient phenotyping from clinical narratives. PLoS ONE 13(2): e0192360.
5. Goldberger, A., Amaral, L., Glass, L., Hausdorff, J., Ivanov, P. C., Mark, R., ... & Stanley, H. E. (2000). PhysioBank, PhysioToolkit, and PhysioNet: Components of a new research resource for complex physiologic signals. Circulation [Online]. 101 (23), pp. e215–e220

3.2. Phenotyping Task 2:

- Can we estimate sample sizes required for building a classifier for phenotype "*Other Substance Abuse*" based on learning curves for related phenotype "*Alcohol Abuse*"?
  - Focused on building a classifier for identifying patients with "*Other Substance Abuse*" based on learning curves estimated using "*Alcohol Abuse*" which had a prevalence of 13%.
  - Estimated learning curves with similar prevalence's of 9,11 and 13%.
  - "*Other Substance Abuse*" has a true prevalence of 9%.
  - Performance from a LASSO using 80-20 training-test split on the "*Other Substance Abuse*" phenotype was compared to estimated performance from our proposed sample size determination methods

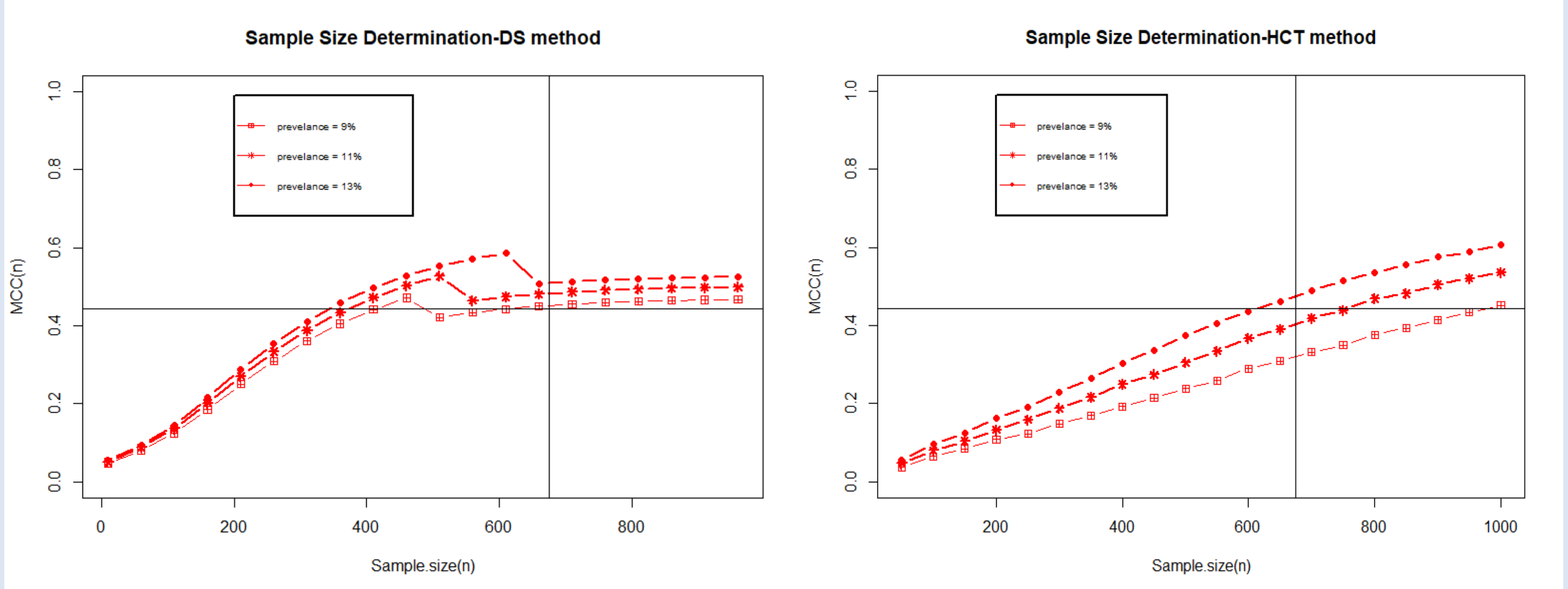
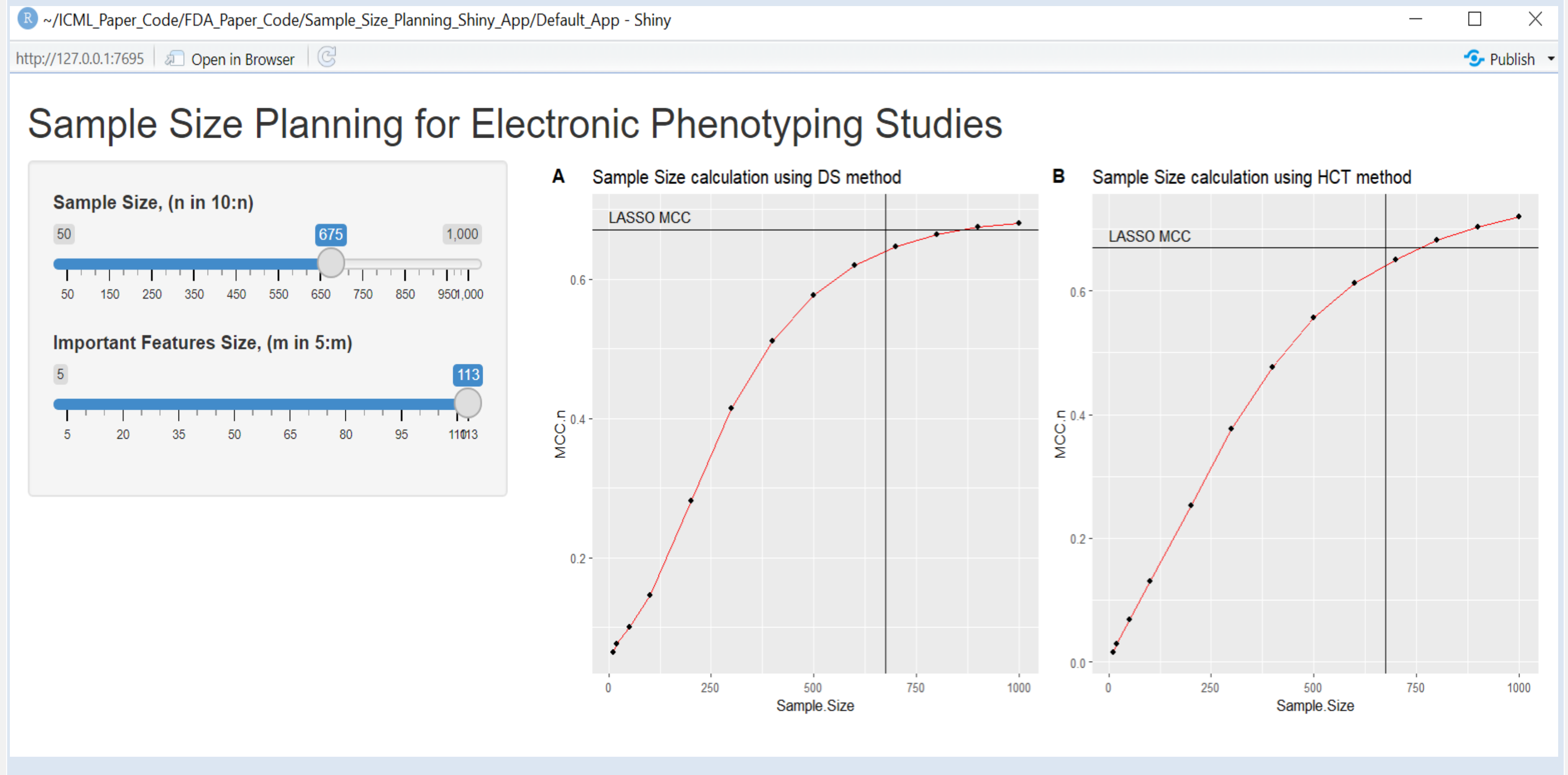


Figure 3: Sample Size determination using DS and HCT for independent features

3.2. Demonstration Using Shiny App:

- We are building a Shiny app for estimating sample sizes required for building a high dimensional LDA classifier using feature selection. The app would work as follows:
  - Inputs : Sample Size (n) Range, Number of Important Features (m) Range, Pilot Data
  - Output : Sample size dependent classification performance curves generated using DS and HCT methods
- A demonstration using a prototype is given below


~/ICML_Paper_Code/FDA_Paper_Code/Sample_Size_Planning_Shiny_App/Default_App - Shiny

http://127.0.0.1:7695

Sample Size Planning for Electronic Phenotyping Studies

Sample Size, (n in 10:n): 675 (range 50 – 1,000)

Important Features Size, (m in 5:m): 113 (range 5 – 113)

Figure 4: Demonstration of Shiny App

6. Acknowledgements

This research was supported by the FDA CDER *Safety Research Interest Group (SRIG)* program and by an appointment to the Research Participation Program at CDER administered by Oak Ridge Institute for Science and Education (ORISE) through an interagency agreement between Department of Energy and FDA.

7. Disclaimer

This presentation reflects the views of the authors and should not be construed to represent FDA's views or policies.