A Resampling-based Ensemble Tree Method to Identify Patient Subgroups with Enhanced Treatment Effect

Chakib Battioui[*†], Lei Shen[*], and Stephen J. Ruberg[*]
*: Lilly Research Laboratories, Eli Lilly and Company, Indianapolis, Indiana

Abstract
In this paper we describe an approach to identify patient subgroups with enhanced treatment effect in clinical trials. It utilizes ensemble trees based on resampling and naturally produces two consistency measures for each potential subgroup identified. We compare simple ways to combine these measures into an overall summary of strength. Using stratified permutations and out-of-bag samples, the approach also provides a multiplicity-adjusted p-value and bias-corrected estimate of treatment effect, both of which are important for decision-making in tailored therapeutics applications. A simulation study is performed to evaluate the performance of the proposed method.

Key Words: Tailored Therapeutics, Predictive Biomarker, Multiplicity

1. Background

In randomized clinical trials, individuals are assigned randomly to a treatment group and a control group. Efficacy and safety outcomes are measured and compared between the two groups. The main interest of the study investigators is to evaluate the treatment effect on the overall population. However, some subgroups of patients may have greater response to treatment than the overall population. It has been well known that patients respond to drugs differently, with many factors that affect the response to any given drug such as genetic makeup, phenotypic, pharmacokinetic, social, and disease severity as well as demographic factors. Increasingly in pharmaceutical drug development, it is not enough to merely show the mean effect of a new treatment is statistically significantly better than the control. Patients, physicians, and payers want and, in fact, are demanding to know more about individual patient outcomes[1], so that the right drug can be selected to properly fit each patient. It has therefore become important to improve on the traditional "one size fits all" paradigm of drug development, and there are now examples of marketed compounds that make tailored therapeutics a reality, such as trastuzumab (Herceptin), imatinib (Gleevec), and cetuximab (Erbitux)[1].

This challenge of identifying subgroups of patients with more desirable clinical outcomes has also been a complex problem for statisticians. Traditional subgroup analyses are based on interaction tests where differential treatment effects among subgroups are analyzed by testing treatment by subgroup interactions in regression models. Such analyses have many drawbacks, such as the inability to consider more complex subgroups involving multiple markers. These limitations have led to many recommendations and generated much caution on the interpretation of results. Many researchers proposed that subgroup analysis should be (1) limited to a few clinically important questions proposed in advance; (2) based on formal tests of interaction; (3)

†: Contact: *Battioui_Chakib@Lilly.com*

adjusted for multiplicity; and (4) fully reported (including all analyses performed) and not over-interpreted[2-4]. However, inappropriate analyses continue to appear in the literature, and there have been many examples of apparently important findings on treatment effect heterogeneity that are subsequently shown to be false[4].

Recently, a number of approaches to subgroup identification have been proposed[5-10] that utilize more advanced statistical methodologies. Two of the techniques found in many of these approaches are recursive partitioning and resampling. In this paper, we propose a rigorous and sophisticated approach to apply these techniques in order to identify subgroups with enhanced treatment effects with controlled type I error rate and improved power.

2. Method

Our method uses recursive partitioning, which has been shown to be extremely useful in modern data mining problems thanks to many attractive features including minimal assumptions on distributions and models[11]. Furthermore, the fact that it directly leads to patient subgroups—as opposed to regression models from some other types of analyses—closely matches the needs for drug development. For tailored therapeutics, a subgroup definition is required to enable the design of a subsequent trial, labeling of the drug by regulatory agencies, and medical decision making by prescribers.

A single analysis of recursive partitioning on a given dataset may not be very stable, as small change in the dataset can lead to quite different results. Significant improvements can be made in this regard by using an ensemble approach enabled by resampling techniques such as bootstrap, cross-validation, and subsampling[12-14]. Although these are similar, subsampling (that is, sampling without replacement) is preferable since it provides the most flexibility in terms of the number and dimensions of the resampled datasets. The same recursive partitioning analysis is performed for each resampled dataset, and we further enrich the ensemble of candidate subgroups by harvesting multiple subgroups from a given tree and consider multiple competing trees for each subsample dataset. By aggregating similar subgroups identified in this ensemble approach, we can easily summarize the frequency by which a given subgroup is identified among resampled datasets, which provides a highly robust measure that we will refer to as “internal consistency”.

An additional benefit of subsampling that we also take advantage of is the out-of-bag sample consisting of observations not included in a given resampled dataset. Since these data are entirely distinct from the corresponding resampled dataset, they can be used to assess, in an unbiased manner, any subgroup findings from the subsample dataset. Similar to the internal consistency, the results of this assessment can be averaged across subsampled datasets pairs (of in-bag and out-of-bag samples) to yield an “external consistency”. While more complex choices can be made, we have found it both simple and useful to assess a subgroup finding from an in-bag sample by determining whether it is directionally consistent in the corresponding out-of-bag sample, and calculating the percentage of consistent out-of-bag samples among all the times this potential subgroup was selected in in-bag samples.

Once we have obtained both the internal and external consistency measures, which contain distinct and complementary information, it is natural to ask how we can best combine the two in order to measure the strength of an identified subgroup. For the remainder of this paper, we will use M_i ("i" for "internal") and M_e ("e" for "external") to denote these two measures for a given subgroup finding. Perhaps the most intuitive and obvious choices for combining the two are: $\min(M_i, M_e)$ and $M_i \times M_e$. The rationale for the first is to require a subgroup to have a minimum level of both internal and external consistencies. It should be more beneficial than using M_i or M_e alone, provided that they manifest on comparable scales, so that one is not always greater than the other. The second combination is the product of the two measures, which would always utilize information contained in both measures. There are of course many other reasonable ways to combine the two measures—for example we can weigh the two unequally—but it will be more complex to investigate those, which is an interesting area of future research.

As an initial investigation of the performance of different measures, we performed a simulation study using datasets with 240 subjects and 20 markers, including one marker that defines a subgroup with enhanced treatment effect. Recursive partitioning with subsampling as described above yielded a number of potential subgroups for each datasets, all with various M_i and M_e values. Four overall consistency summary for these subgroups were considered: (1) M_i alone (2) M_e alone (3) $\min(M_i, M_e)$, and (4) $M_i \times M_e$. For each summary, the type I error rate and power were estimated for each possible "critical value" of the summary by calculating the respective numbers of correct and incorrect subgroups whose summary exceeded the critical value. The power curves (power vs. type I error rate) for the four summaries (Figure 1) demonstrate the superiority of the last summary, $M_i \times M_e$, in this setting.

When it is desirable to produce a multiplicity-adjusted p-value for the strongest subgroup identified, our method utilizes permutation that is stratified by treatment groups. That is, a permuted dataset is obtained by shuffling the observed responses within each treatment arm. While all permutation methods require (often implicitly) assumptions to be valid, this specific permutation scheme preserves the overall treatment effect and is an ideal match to the tree construction method (that is, trees are constructed first from one of the two treatment arms), hence is expected to be quite robust. As is standard, after performing a large number of simulations, the summary of the best subgroup identified from each permuted dataset provides a reference distribution, with which the summary of the top subgroup from the actual data is compared to yield a multiplicity-adjusted p-value.

Besides the external consistency measure described above, an additional benefit provided by the out-of-bag samples is an unbiased estimate of the differential treatment effect associated with the subgroup. It is well known that the "naïve" estimate of the size of an effect from the same data that led to the identification of this effect is upwardly biased, sometimes severely so. In the data mining literature, the best option to replicate a finding, including the size of the effect, is to utilize an independent dataset. Put in the drug development context, this means either a new clinical study, or having sufficient amount of data from the current clinical study so that part of that data is set aside as "testing data" to be used, not in the identification of subgroups, but only in validation of an identified subgroup. However, given the fact that a clinical study is typically powered to detect a main treatment effect, coupled with the lower power of detecting a treatment by subgroup interaction, not surprisingly in practice it is rare to have the luxury of a sufficiently large study to enable the setting aside of a testing dataset. In such situations,

the out-of-bag samples made possible by bootstrap or subsampling provide the next best solution to obtaining an unbiased estimate of the differential treatment effect.

The proposed method can therefore be described by the following algorithm:

1. Sample the original data *B* times (done separately for each treatment arm), each time creating a pair of mutually exclusive datasets (in-bag and out-of-bag samples) with size as specified percentages (such as 50%-50% or 70%-30%) of the original dataset.

2. Harvest potential subgroups of enhanced treatment effect for each in-bag dataset by first building a tree with a specified maximum depth using a specified treatment arm, and then combining with the other treatment arm and applying specified selection criteria (for example the observed treatment effect in the subgroup needs to be enhanced beyond a certain threshold as compared to the observed overall treatment effect). The number of potential subgroups identified from each resampled dataset also depends on the specified number of competing markers to be considered; for example if one competing marker is considered, then after the strongest subgroup is identified the analysis is re-run without the corresponding marker. The purpose of considering competing markers is to avoid "masking" of markers and subgroups.

3. Each identified subgroup is assessed for consistency and differential treatment effect in the corresponding out-of-bag sample.

4. Combining results across subsampled dataset pairs, the internal and external consistency measures are calculated for each identified subgroup. The two can be combined (we use the product $M_i \times M_e$) to produce an overall summary.

5. Using permutation stratified by treatment arms, a large number of permuted datasets are obtained, each analyzed as described above. This provides a reference distribution of the summary measure, against which the observed results from the actual dataset is compared to yield a multiplicity-adjusted p-value.

We have implemented this method using SAS (SAS 9.2, Enterprise Miner 6.1 and specifically Proc Arbor) and R. A key consideration in practice is the computing speed, and to that end, the architecture of the method lends itself naturally to parallel computing that can dramatically improve the speed if a large number of computing nodes are utilized.

3. Simulation Study

A simulation study was performed to assess the proposed method. Each generated dataset consists of a number (represented by *p*) of 3-level genetic markers (with values 0, 1, and 2, representing the number of miner alleles a subject is carrying for a given SNP), a continuous outcome *Y*, and a binary treatment variable *T* (representing "Treated" and "Placebo" groups). The outcome *Y* was generated from a linear model, where the mean placebo response is -0.1, and the standard deviation conditional on all markers is 1.13. The number of markers (*p*) can be 5, 20, or 50 and the sample size (*n*) is either 240 or 480. In terms of marker effects, datasets were generated under both the "null" scenario (that is, no predictive marker) for evaluation of type 1 error rate and "alternative"

scenarios with one or two predictive biomarkers for assessment of statistical power. When predictive biomarkers are present, the mean treatment effect in the weakest-responding subgroup is -0.1, and each predictive marker is associated with a differential treatment effect of -0.45. A subgroup is considered to be identified if the multiplicity-adjusted p-value is less than 0.1. Results of the simulations are presented in Tables 1-7 and Figures 1-3.
Table 1 compares the performance under the null scenario when the number of markers ranges from 5 to 50 (scenarios A, B, C), and the results show that type I errors are controlled at close to the nominal level.

Results for scenarios D, E, and F are presented in Table 2 and Figure 2. Here datasets were generated with 1 predictive marker, and the number of markers again ranges from 5 to 20 to 50. The overall statistical power is low, in that no subgroup was identified for 55%-74% of datasets across the scenarios. However, when looking at the instances when at least one subgroup is identified, the performance of the method is good, especially when the number of markers is small. In other words, when a subgroup is identified it tends to be a correct subgroup. Table 3 provides additional performance summaries for these three scenarios at the subgroup level.

To assess the performance under different combinations of p and number of predictive markers, in Table 4 we summarized results when the total number of markers is 20 or 50, and the number of predictive markers is 1 or 2. Comparing scenarios E and G, we can see that the conditional power of identifying both predictive markers is about half of that of identifying the lone predictive marker, when there are 20 markers overall. The drop is smaller when there are 50 markers overall. Table 5 provides additional subgroup-level performance summaries.

The impact of sample size is illustrated in Tables 6-7 and Figure 3, where for each p (20 or 50) and predictive markers (1 or 2), sample sizes of 240 and 480 are compared. Large gain of statistical power is seen across the board.

4. Conclusions

We have described a resampling-based ensemble tree approach to identify subgroups of patients with enhanced treatment effect in clinical trials. It has a number of advantages:

- The recursive partitioning approach determines subgroups, a good match with drug development and medical and regulatory decision making;

- By using an ensemble approach, the results are robust to outliers, which reduces spurious findings to which some other methods are prone;

- Criteria such as minimum subgroup size can be applied to eliminate subgroups that do not meet the need of a specific project, thus reducing the scope of the overall “search space” and lessons the severity of multiplicity, leading to increase statistical power;

- By allowing a specified number of competing markers in the "harvesting" of trees, the issue of collinearity is easily handled, so that a potentially useful marker is not masked by others;

- The out-of-bag samples conveniently supplied by bootstrap or subsampling provide key information such as directional consistency and bias-corrected estimate of effect;

- By intelligently utilizing both the internal and external consistency measures, the power is improved for a given level of type I error rate control.

Furthermore, although we have primarily dealt with the more challenging problem of identifying super-responder subgroup identification that is common in tailored therapeutics, the same approach can be used to identify prognostic factors from a single-arm clinical trial.

Because of the need to perform nested resampling (for example permutation and subsampling), this approach can be computationally intensive. However, the architecture of the approach lends itself naturally to parallel computing, which can be leveraged to dramatically improve the computing speed.

There are a number of interesting areas for further research, such as optimization of how the internal and external consistency measures can be combined. It would also be informative to evaluate the performance of the method in additional scenarios.

Acknowledgement

The authors would like to thank colleagues Hollins Showalter and Brian Denton for their help, especially with the simulation study.

References

1. Ruberg SJ, Chen L, Wang Y (2010) *The mean does not mean as much anymore: Finding subgroups for tailored therapeutics.* Clinical Trials 7: 574–583

2. Brookes ST, Whitley E, Peters TJ, Mulheran PA, Egger M, Davey Smith G (2001) *Subgroup analyses in randomised controlled trials: quantifying the risks of false-positives and false-negatives.* Health Technol Assess. 5(33):1-56

3. Wang R, Lagakos SW, Ware JH et al. (2007*) Statistics in medicine – reporting of subgroup analyses in clinical trials.* New Eng J Med 357: 2189–94

4. Rothwell PM (2005) *Subgroup analysis in randomized controlled trials: importance, indications, and interpretation.* Lancet 365: 176–86

5. Negassa A, Ciampi A, Abrahamowicz M, Shapiro S, Boivin J-F (2005) *Tree-structured subgroup analysis for censored survival data: Validation of computationally inexpensive model selection criteria.* Statistics and Computing 15:231–239

6. Su XG, Zhou T, Yan X, Fan J, Yang S (2008) *Interaction trees with censored survival data.* International Journal of Biostatistics 4(1) Article 2

7. Su X, Tsai CL, Wang H, Nickerson DM, Li B (2009) *Subgroup analysis via recursive partitioning.* Journal of Machine Learning Research 10:141–158

8. Lipkovich I, Dmitrienko A, Denne J, Enas G (2011) *Subgroup identification based on differential effect search- A recursive partitioning method for establishing response to treatment in patient subpopulations.* Statistics in Medicine 30(21):2601-21

9. Foster JC, Taylor JM, Ruberg SJ (2011) *Subgroup identification from randomized clinical trial data.* Statistics in Medicine 30(24):2867-80

10. Loh, W-Y, Man, M, He, X (2013) *A regression tree approach to subgroup identification for censored data*. Presented at Joint Statistical Meetings 2013 and submitted for publication

11. Breiman L, Stone CJ (1984) *Classification and Regression Trees.* New York: Chapman & Hall

12. Breiman L (1996) *Bagging markers.* Machine Learning 24:123-140

13. Breiman L (2001) *Random Forests.* Machine Learning 45:5–32

14. Friedman J (2001) *Greedy function approximation: A gradient boosting machine.* Annals of Statistics 29(5):1189-1232

Table1: Estimated type I error rate for identifying predictive markers (n=240)

Scenario	Markers	Type I Error Rate
A	5	0.10
B	20	0.11
C	50	0.12

Table 2: Summaries of estimated power for identifying predictive markers (n=240, 1 predictive marker)

Scenario	Markers	No Subgroup Identified	(Conditional) Sensitivity	(Conditional) Specificity	(Conditional) PPV	(Conditional) NPV
D	5	55%	0.955	0.988	0.955	0.988
E	20	62%	0.769	0.986	0.763	0.988
F	50	74%	0.538	0.989	0.519	0.990

Table 3: Summaries of identified subgroups (n=240, 1 predictive marker)

Scenario	Markers	No Subgroup Identified	Average Size of Subgroup	Average Treatment Effect
D	5	55%	0.500	-0.540
E	20	62%	0.521	-0.495
F	50	74%	0.535	-0.452

Table 4: Summaries of estimated power for identifying predictive markers (n=240)

Scenario	Markers	Predictive Markers	No Subgroup Identified	(Cond.) Sensitivity	(Cond.) Specificity	(Cond.) PPV	(Cond.) NPV
E	20	1	62%	0.769	0.986	0.763	0.988
G	20	2	52%	0.395	0.988	0.791	0.936
F	50	1	74%	0.538	0.989	0.519	0.990
H	50	2	60%	0.462	0.996	0.850	0.978

Table 5: Summaries of identified subgroups (n=240)

Scenario	Markers	Predictive Markers	No Subgroup Identified	Average Size of Subgroup	Average Treatment Effect
E	20	1	62%	0.500	-0.540
G	20	2	52%	0.528	-0.737
F	50	1	74%	0.535	-0.452
H	50	2	60%	0.518	-0.752

Table 6: Summaries of estimated power for identifying predictive markers

Scenario	Markers	Predictive Marker	Sample Size	No Subgroup Identified	(Cond.) Sensitivity	(Cond.) Specificity	(Cond.) PPV	(Cond.) NPV
E	20	1	240	62%	0.769	0.986	0.763	0.988
I	20	1	480	37%	0.968	0.994	0.936	0.998
F	50	1	240	74%	0.538	0.989	0.519	0.990
J	50	1	480	49%	0.941	0.997	0.921	0.998
G	20	2	240	52%	0.395	0.988	0.791	0.936
K	20	2	480	15%	0.688	0.996	0.958	0.967
H	50	2	240	60%	0.462	0.996	0.850	0.978
L	50	2	480	24%	0.638	0.998	0.956	0.985

Table 7: Summaries of identified subgroups

Scenario	Markers	Predictive Marker	Sample Size	No Subgroup Identified	Average Size of Subgroup	Average Treatment Effect
E	20	1	240	62%	0.500	-0.540
I	20	1	480	37%	0.522	-0.539
F	50	1	240	74%	0.535	-0.452
J	50	1	480	49%	0.510	-0.536
G	20	2	240	52%	0.528	-0.737
K	20	2	480	15%	0.508	-0.772
H	50	2	240	60%	0.518	-0.752
L	50	2	480	24%	0.516	-0.779

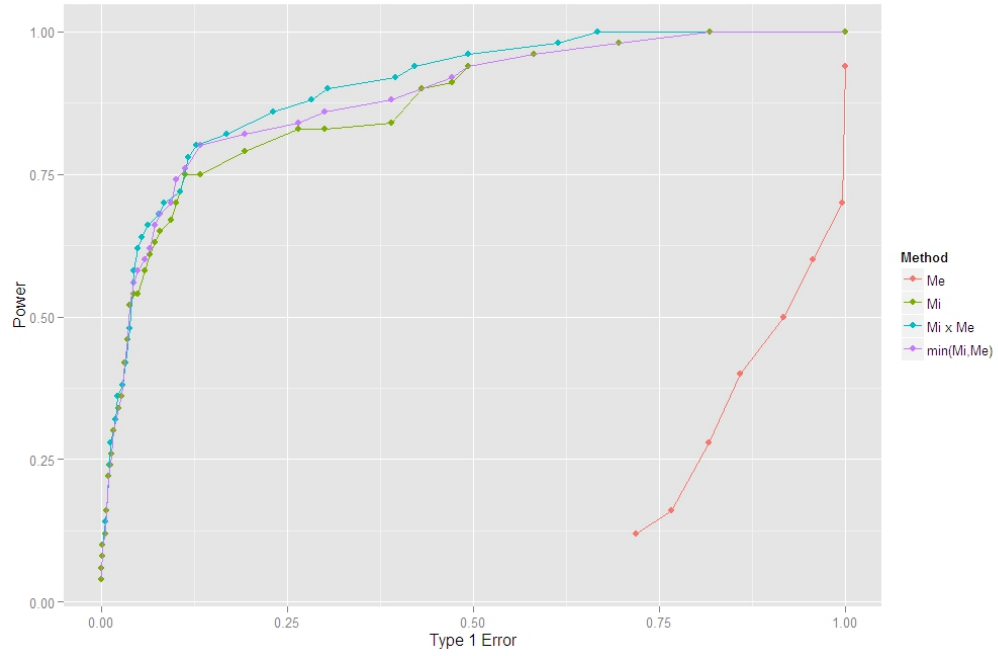

Figure 1: Comparing Performance of Different "Summaries of Strength"

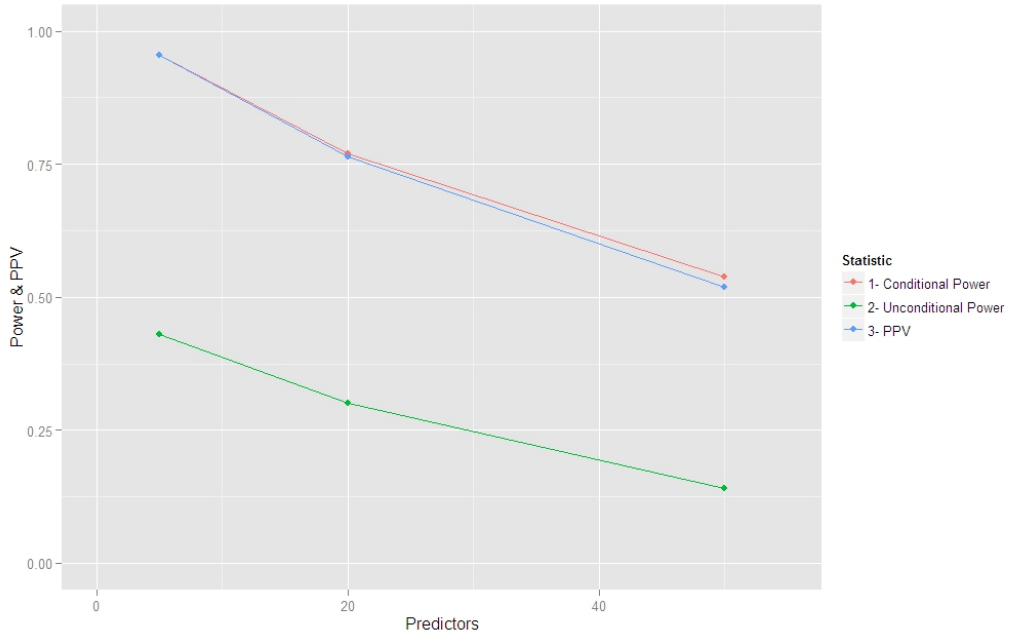

Figure 2: Relationship between Power and Total Number of Markers

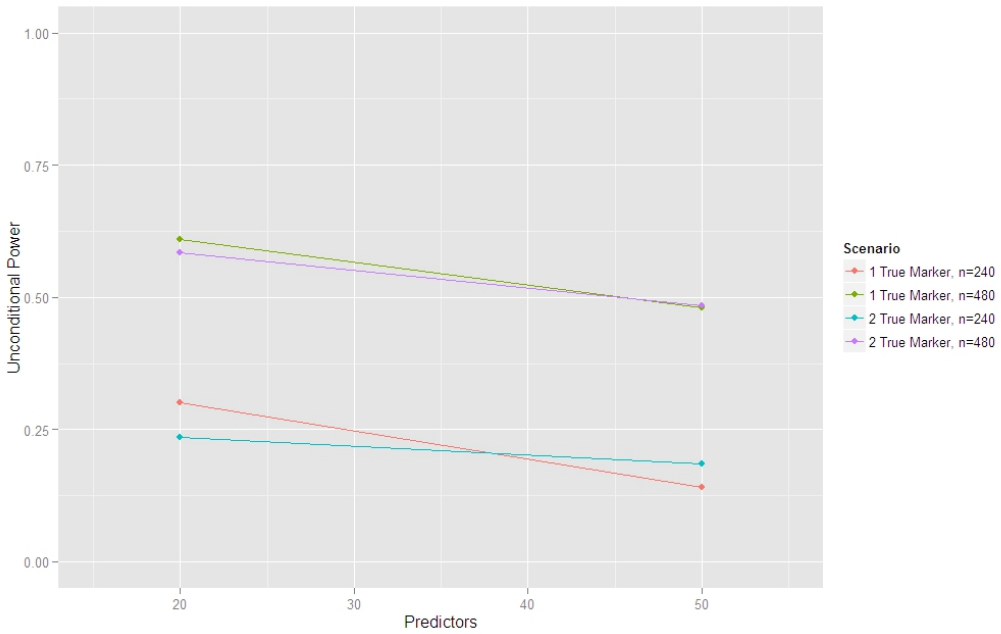

Figure 3: Power Comparison between Scenarios