**Feature Reduction through Imaginary Network Models**

4/16/20

Farrokh Alemi, Ph.D.

George Mason University Department of Health Administration and Policy

[falemi@gmu.edu](mailto:falemi@gmu.edu)

703-893 3799

Sankeerthi Mummidisetty [smummidi@gmu.edu](mailto:smummidi@gmu.edu)

Amr Hassan Elrafey

**ABSTRACT**

**Background:** The concepts introduced in network models, in particular blocking backdoor paths, provide a method of thinking about irrelevant features. Backdoor paths refer to correlated pairs of variables that start from outcome and end up with a variable that both precedes, and is associated with, treatment. A set of variables that blocks all backdoor paths is referred to as Markov blanket. Parents refer to variables in Markov blanket of a target variable that precedes the target variable. In evaluating treatment effectiveness, the causal network theory demonstrates that retaining parents of treatment (PT) or parents of outcome (PO) is sufficient. All other variables can be ignored. **Objective:** This paper demonstrates how to identify PT or PO variables, without creating network models. We also show the loss of accuracy when analysis is limited to PT and PO variables. **Data Source**: In the first example, we examined survival (outcome variable) from stomach (treatment variable) using 35 most common comorbidities (covariates) in data obtained on veterans. In the second example, we examined impact of citalopram (treatment) on remission of depression symptoms (outcome), statistically controlling for 32 baseline comorbidities of the patient. The accuracy of feature reduction was compared to a comprehensive regression model that included all comorbidities. **Principal Findings:** In analyzing survival from stomach cancer, PT method reduced the number of covariates stratified by 71%; PO by 69%. Despite these reductions, the reduced models had same cross-validated accuracy as a model with no reduction, respectively 80.14%, 80.01%, and 81.16%. In analysis of citalopram, the PT reduced the features by PT reduced the features by 71%(35 to 10), PO by 54%(35 to16), and the cross-validated percent of variation explained was 61%,61% and 57% for PT, PO, and no-reduction model. **Conclusions**: Multivariate analysts can use LASSO regression to identify PT and PO variables. Using either of these two methods, the analysts can reduce features used in the model without loss of accuracy.

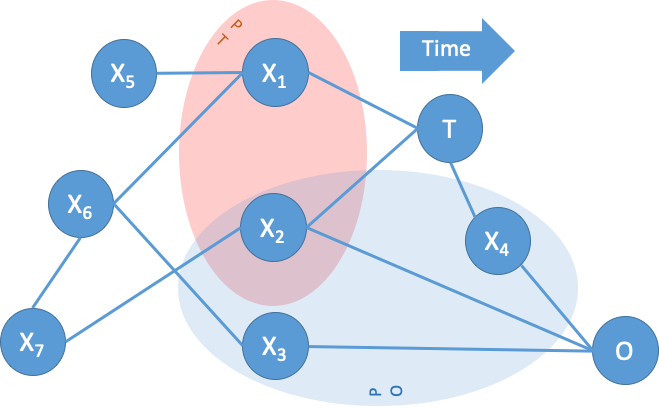
**Keywords**: Feature reduction, Stratification, Covariate Balancing, Markov Blanket, Blocking Backdoors, Probability Network Models, Bayesian Networks, causal networks

**INTRODUCTION**

The concepts introduced by causal networks can be used to reduce features in high dimensional data. This paper provides a non-mathematical introduction to these concepts and then show the implication of this theory for feature reduction. Finally, by way of two examples, we empirically test the theoretically suggested feature reductions.

**A Mathematical Theory**

**Figure 1: Parents of Treatment (T) and Outcome (O)**



Pearl uses a-cyclical, directed, network models to capture relationships among interrelated variables (see Figure 1 for an example). In contrast to regression models, these causal networks consider the timing of variables, as causes must always occur prior to effects. In Figure 1, a movement from left to right indicates next time periods. Causal models also ignore spurious correlations that might exist in the data. Spurious correlations refer to associations that are not stable and disappear when the model is examined in subsets of the data. In Figure 1, the treatment, T, and outcome, O, are shown as well a host of other variables shown as Xs. If a variable is correlated with both the treatment and the outcome variable, then it is considered a covariate; and a potential source of confounding in measuring the impact of treatment on the outcome. A variable is a covariate if starting from the variable, one can follow the links (whether in the direction of the arrows or against the direction) and reach both treatment and outcome variables. Thus, X1 is a covariate because we can reach both T and O by following the links in the network. X1 directly affects T but indirectly is correlated with X6, which is correlated with X3, which has an effect on O. X5 is not a covariate because it does not affect both T and O. It only reaches, or is correlated with, T. If a variable is not a covariate, then feature reduction can ignore the variable with no concerns. This variable does not distort the effect of treatment on outcome and ignoring it has no effect of treatment on outcome.

In contrast, if the variable is a covariate, feature reduction techniques can only ignore it if all “backdoor” paths are “blocked.” A backdoor path starts from outcome and follows, in or against the direction of the links, to a variable with a directed link to treatment (Maathius and ColoPO 2015). For example, starting with O, we can go to X6, then to X3 and finally end with X1, which is both correlated with T and precedes it. In a correlation matrix, and without use of the network model, one can identify a backdoor path by identifying a set of variables in which overlapping pairs of variables are correlated with each other, the first pair is a variable correlated with outcome, and the last pair is a variable correlated with and preceding treatment. For our previous paths, we start with the correlation of O and X6, then correlation of X6 and X3, and finally end with correlation of X1 and T and the fact that X1 precedes T. Even though the analyst does not know the network model, he/she can identify the backdoor paths easily from the correlation matrix.

A backdoor path can be blocked if any variable in the path is “stratified.” Stratification, or in probabilistic terminology “conditioning”, refers to re-examining the relationships among the variables in subgroups of the data. For example, stratification of the binary variable X1 refers to examining the relationships among the variables under two conditions: first when X1=1 and then when X1=0. The net effect across strata can be assessed and this effect is typically different than the average effect without stratification. Pearl’s analytical studies have shown that blocking *all* the backdoor paths is sufficient and necessary condition for calculating the unbiased impact of treatment on outcome. A set of variables that block all backdoor paths to a variable is called “Markov blanket” of the variable. In Figure 1, the Markov blanket of T includes X1, X2, and X4. A Markov blanket splits the network into two. All variables outside the blanket can be ignored with no loss of accuracy. Empirical research also supports this finding: a number of investigators have demonstrated that Markov blankets of a target variable can radically reduce features needed to predict it (Fu et al. 2010; Yifeng 2009; Tan and Zhifa 2013; Shen et al. 2008; Zeng et al 2009, Aliferis et al. 2003).

Beyond ignoring the variables outside the Markov blanket, there are also some variables inside the Markov blanket that can, and should, be ignored. Within a Markov blanket, one can separate the variables into “parents, children, and co-parents”. Parents refer to variables that precede the target variable (X1 is a parent to T in Figure 1), meaning that they occur before the target variable. Children refer to variables that follow the target variable (X4 is a child of T). Co-parents refer to parents of the children of a target variable (In Figure 1, X5 and X6 are co-parents of X1). Children of treatment should never be stratified because they could be on the causal path to outcome and their stratification will distort the relationship between treatment and outcome. This allows us to focus on parents within Markov blanket and ignore all other variables. In Figure 1, parents of T are X1 and X2 and are shown within an oval marking them as PT. Parents of O are X2, X3 and X4. Note that X4 is a child of treatment and should not be retained.

**Methods of Identifying Parents & Children**

Several methods exist for finding parents in the Markov Blanket of a target variable. These methods include the Grow-Shrink algorithm (Aliferis et al. 2003) or the Partial Correlation method (Opgen-Rhein and Strimmer 2007). The method that we prefer is based on LASSO regression, a procedure familiar to multivariate analyst. Li, Dai and Tu (2004) show that LASSO regressions identify the variables in the Markov blankets. Shojaie and Michailidis (2010) show that parents in Markov blanket can be identified through examining timing of events. Given a sequentially ordered set of variables, they used LASSO regression to identify the parents in the Markov Blanket. Since our goal is to block all backdoors, there are at least two ways to proceed using the parents in the Markov Blanket of either treatment or outcome:

1. If outcome is the last variable measured, then all backdoor paths can be blocked by stratifying parents in the Markov Blanket of outcome (PO), if no child of treatment is included. The network theory, informs us that we should block all backdoor paths from outcome to treatment. This can be done by retaining treatment, not children of treatment, and all other parents of outcome. Blocking these paths will be sufficient to block all paths. No covariates will remain that affect both the outcome and treatment. If indicates the regression parameters, is a constant that restricts the number of non-zero significant variables in the model, then the LASSO regression equation for this situation is given by the following formula:

Note that in this regression the response variable is the outcome, independent variables are treatment and covariates that occur before treatment (no children of treatment are included). LASSO regression selects variables that if statistically controlled should remove the effects of all remaining variables in the regression model.

1. The backdoor path can also be blocked by stratifying the parents to exposure/treatment (PT). By definition, all backdoor paths end with a variable that is parent to exposure/treatment. The causal network theory claims that if the parents of treatment are stratified then all backdoor paths are blocked. No covariates remain that directly affect treatment. The LASSO regression equation for this situation is given by the following formula:

Note that the response variable in this regression is the treatment, and not the outcome, variable. All variables that occur after treatment, including outcome, are excluded from the analysis. All independent variables occur prior to treatment. LASSO regression selects a set of variables that if stratified will remove the effects all other covariates occurring prior to treatment.

**Methods for First Data Set**

**Source of Data:** Data were obtained from Veterans Administrations data warehouse. We examined survival time of veterans in 168 Medical Centers over seven years starting from 1/2008 to 12/2015. To be included in the cohort, patients had to have 2 primary care visits, not more than 2 years apart, to a VA facility. 4,681,809 veterans had at least two primary care visits. Among these we focused on 829,827 veterans who were hospitalized.

**Outcome:** The primary outcome was survival within 6 months of the date of the first diagnoses of stomach cancer.

**Treatment/Exposure:**  In this study, treatment/exposure variable is cancer of stomach.

**Covariates:** The independent variables were comorbidities of stomach cancer. There were 10,292 unique diagnoses in the hospitalization data. We excluded from the analysis any diagnoses that occurred after stomach cancer, as these were considered possible complications of cancer and potentially on the causal path from cancer to survival. We also excluded any diagnoses that did not occur at least 100 times in patients who had stomach cancer. The common ICD 9 codes that were included in the analysis as potential covariates of the cancer were:

|  |  |
| --- | --- |
| 403.90 Hypertensive renal disease  427.31 Atrial fibrillation  428.0 Congestive heart failure  600.00 Benign hypertrophy of prostate without urinary obstruction  585.9 Chronic kidney disease  414.00 Coronary atherosclerosis  599.0 Other disorders of urinary tract  414.01 Coronary atherosclerosis  244.9 Hypothyroidism  V66.7 Encounter for palliative care  584.9 Acute kidney failure  578.9 Hemorrhage of gastrointestinal tract  V58.61 Long term (current) use of anticoagulants  197.7 Malignant neoplasm of liver  309.81 Posttraumatic stress disorder  305.1 Tobacco use disorder | 486. Pneumonia  458.9 Hypotension  496. Chronic airway obstruction  285.9 Anemia  280.9 Iron deficiency anemia  272.4 Other and unspecified hyperlipidemia  564.00 Constipation  787.20 Dysphagia  401.9 Unspecified essential hypertension  511.9 Unspecified pleural effusion  V15.82 Personal history of nicotine dependence  276.51 Dehydration  263.9 Unspecified protein-calorie malnutrition  530.81 Esophageal reflux  276.8 Hypopotassemia  276.1 Hypo-osmolality and hyponatremia  150.9 Malignant neoplasm of esophagus  E849.7 Unspecified place residential institution as the place of occurrence of the external cause  311. Depressive disorder |

These covariates could radically change prognosis of stomach cancer. It is important to measure to what extent these covariates affect both hospitalization for stomach cancer and survival from it.

**Methods of identifying Markov Blanket**: Our preferred method of identifying parents in the Markov Blanket was to use LASSO regression. In LASSO regression, variables are selected by effect size. One way to accomplish LASSO regression within an ordinary logistic regression is to require that the standardized coefficient not be between plus or minus arbitrary cutoff point, set in our study to log(1.5). The use of odds of 1.5 times is arbitrary but follows the practice in statistical process control literature to consider doubling of odds of an event as clinically significant change.

**Methods of Measuring Accuracy:** The accuracy of PT, OT, and comprehensive stratifications is reported using Area under the Receiver Operating Curve. To make sure that we do not model noise, 10-fold cross validation was carried out. 10% of the data were set aside for validation and the remaining data were used to estimate model parameters.

**Results:** When we focused on hospitalization comorbidities that occurred at least 100 times in the data, and occurred prior to cancer, 35 diagnoses were identified. These diagnoses were assumed to be the common comorbidities, and not complications, of stomach cancer. The main-effect LASSO regression of treatment on covariates that precede it is provided in Table 1. The following 8 comorbidities were parents in the Markov Blanket of stomach cancer: 309.81, 150.9, 263.9, 511.9, 787.20, 280.9, 197.7, 578.9. Similarly, the following 8 comorbidities were parents on the Markov Blanket of survival: stomach cancer, 309.81, 150.9, 263.9, 511.9, 787.20, 197.7, V66.7.

**Table 1: Comorbidities with Significant and Large Impact on Cancer or Survival**

(Effect size should be higher than +.4 or lower than -.4 to be considered a large effect)

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| **Inpatient Diagnoses Codes** | **Regression of Stomach Cancer on its Comorbidities** | | |  | **Regression of 6-Month Survival on Stomach Cancer & its Comorbidities** | | |
| **Effect Size** | **P-Value** | **Parents to Exposure/Treatment** |  | **Effect Size** | **P-Value** | **Parents to Outcome** |
| 309.81 | -0.552 | <0.001 | Yes |  | -0.511 | < 0.001 | Yes |
| 150.9 | 2.554 | < 0.001 | Yes |  | 1.484 | < 0.001 | Yes |
| 263.9 | 0.697 | < 0.001 | Yes |  | 0.692 | < 0.001 | Yes |
| 276.51 | 0 | N/A | No |  | 0 | N/A | No |
| 511.9 | 0.431 | < 0.001 | Yes |  | 0.491 | < 0.001 | Yes |
| 787.20 | 0.939 | < 0.001 | Yes |  | 0.544 | < 0.001 | Yes |
| 280.9 | 0.839 | < 0.001 | Yes |  | 0 | N/A | No |
| 197.7 | 2.193 | < 0.001 | Yes |  | 2.793 | < 0.001 | Yes |
| 578.9 | 1.153 | < 0.001 | Yes |  | 0 | N/A | No |
| V66.7 | 0 | N/A | No |  | 3.292 | < 0.001 | Yes |
| 585.9 | 0 | N/A | No |  | 0 | N/A | No |
| 403.90 | 0 | N/A | No |  | 0 | N/A | No |
| 305.1 | 0 | N/A | No |  | 0 | N/A | No |
| E849.7 | 0 | N/A | No |  | 0 | N/A | No |
| 276.1 | 0 | N/A | No |  | 0 | N/A | No |
| 530.81 | 0 | N/A | No |  | 0 | N/A | No |
| V15.82 | 0 | N/A | No |  | 0 | N/A | No |
| 401.9 | 0 | N/A | No |  | 0 | N/A | No |
| 564.00 | 0 | N/A | No |  | 0 | N/A | No |
| 272.4 | 0 | N/A | No |  | 0 | N/A | No |
| 285.9 | 0 | N/A | No |  | 0 | N/A | No |
| 427.31 | 0 | N/A | No |  | 0.496 | < 0.001 | No |

Table 2 shows the estimated impact of stomach cancer on survival in comprehensive, PT and OT methods of controlling for confounding. These data show that PT reduced the number of covariates from 35 to 10 variables, a 71% reduction. PO reduced the number of covariates from 35 to 11 variables, a 69% reduction. Despite these reductions, the accuracy of the three methods was comparable: the 10-fold cross-validated Area under the Receiver Operating Characteristic (AROC) for predicting survival from the selected covariates was 80.14%, 80.01%, and 81.16% for comprehensive, PO and PT methods, respectively. These data suggest that the reduction in variables stratified did not affect the predictive accuracy. Surprisingly, the PT had even higher 10-fold cross-validated AROC than the comprehensive approach, which is possible if the comprehensive approach of stratifying everything leads to modeling noise in the training set.

**Table 2: Impact of Stratification on 6-Month Survival from Stomach Cancer**

(AROC is Area under Receiver Operating Curve in using the selected covariates to predict survival.

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **Variables in the Model** | **Number of Covariates** | **Outcome Odds Ratio (95% Confidence Interval)** | **Number of Cases Matched (%)** | **Number of Controls Matched (%)** | **10-fold Cross-Validated AROC** | **AROC divided by Degrees of Freedom** |
| Comprehensive | 35 | 0.157 (0.135 to 0.183) | 439 (25.3%) | 338,462 (40.88%) | 80.14% | 2.357 |
| Only Parents of Survival, Cancer & No Cancer Children | 11 | 0.221 (0.204 to 0.209) | 1,456  (84.16%) | 746,443 (89.9%) | 80.01% | 8.001 |
| Only Parents to Cancer | 10 | 0.244 (0.227 to 0.262) | 1,532 (88.5%) | 808,788 (97.6%) | 81.16% | 9.018 |

**Results for Second Data Set**

In the second data set, we evaluate the effectiveness of an antidepressant on reducing depression symptoms.

**Source of Data:** We used the National Institute of Mental Health’s STAR\*D data, the largest available data set capturing response to antidepressants. These data include genetic markers and phenotypes of 3,678 patients with major depression. The STAR\*D cohort data was split into the training (2931 patients) and research data (747 patients) sets. The training dataset was derived from approximately 80% of STAR\*D data, while the evaluation dataset was derived from the remainder (~20%) of STAR\*D data. The STAR\*D includes up to 4 episodes of treatment for patients who do not respond to the initial treatment. The unit of analysis was the episodes of treatment; thus, we are comparing patients treated with one antidepressant to patients treated with another antidepressant, what in the literature is known as placebo studies.

**Outcome:** The outcome of interest is depression symptom remission, measured in STAR\*D through 50% reduction in Hamilton’s index of depression symptoms.

**Treatment:**We focus on the effectiveness of citalopram, one of the antidepressants provided to patients in STAR\*D trials.

**Covariates:** The effectiveness of the antidepressant depends in part on gender and patient’s history of illness, captured by baseline comorbidities, listed below:

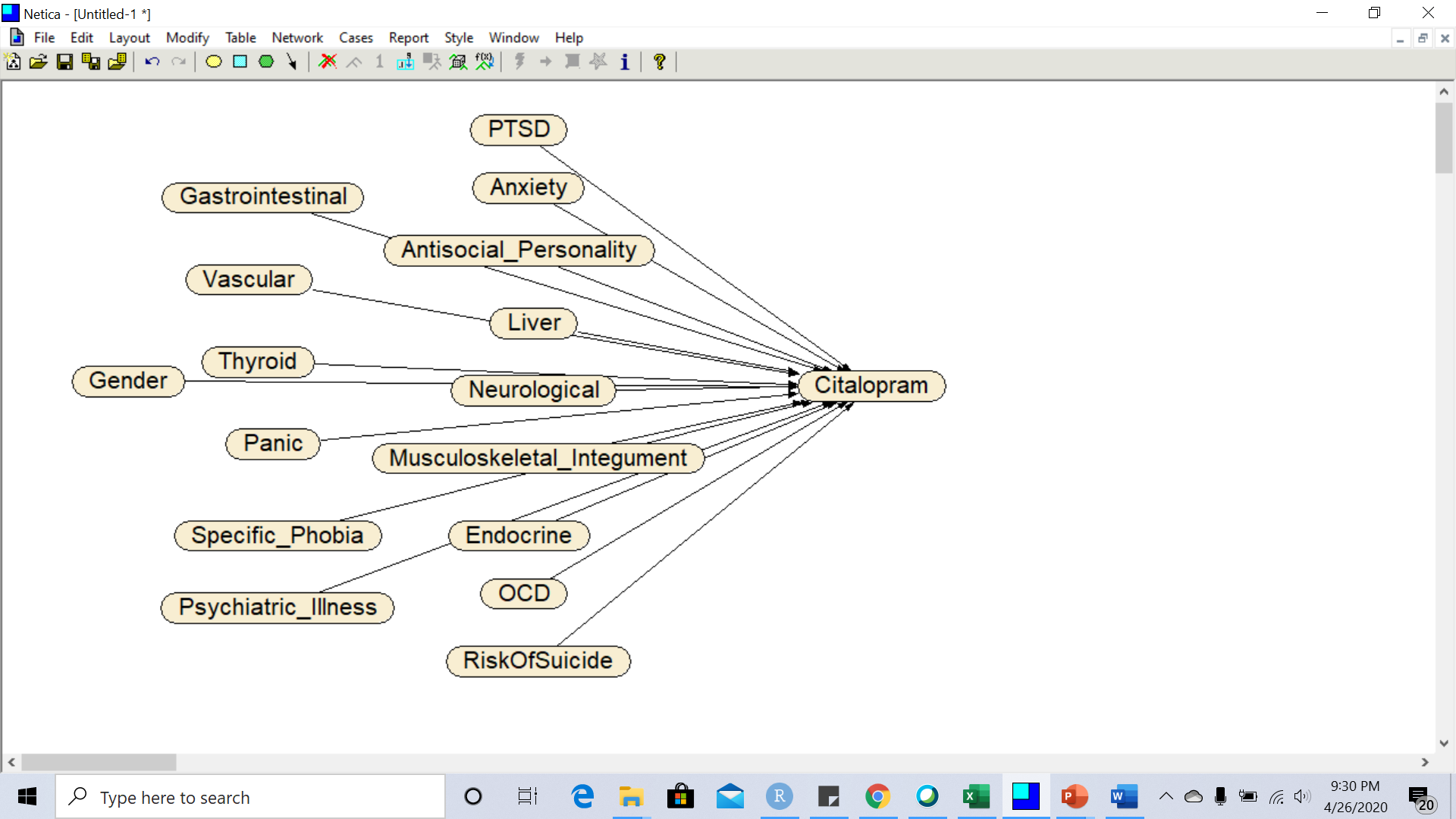
|  |  |
| --- | --- |
| Risk of Suicide | Thyroid |
| Vascular | Heart |
| Eyes Ears Nose Throat Larynx | Hematopoietic |
| Renal | Gastrointestinal |
| Musculoskeletal Integument | Genitourinary |
| Psychiatric Illness | Neurological |
| Liver | Respiratory |
| Alcohol | Endocrine |
| Cannabis | Amphetamine |
| Panic | Opioid |
| Social Phobia | Specific Phobia |
| PTSD | OCD |
| Borderline Personality | Anxiety |
| Antisocial Personality | Dependent Personality |
| Personality Disorder | Paranoid Personality |
| Bulimia | Anorexia |
|  | Cocaine |

**Methods of Identifying Markov Blanket**: We used the glmnet R software to conduct the LASSO regression. Demographic variables were assumed to have occurred prior to baseline comorbidities and baseline comorbidities were assumed to occur prior to treatment with citalopram. The initial LASSO regression was treatment on variables that occur prior to it. In the second LASSO regression, the response variable was the outcome, i.e. symptom remission. The independent variables were the variables that were significant in the initial LASSO regression plus the treatment variable.

**Methods of Measuring Accuracy:** We have evaluated the performance of the model using coefficient of determination (*r*2 ) and Root Mean Squared Error (RMSE). R squared is the proportion of the variance in the dependent variable that is predictable from the independent variables Ideally, lower RMSE indicates a better model. In our study we have observed that the RMSE on the training data are 0.465. The results on the test data is 0.467 when considering citalopram as target variable and RMSE on the training data are 0.483. The results on the test data is 0.486 when considering remission as target variable

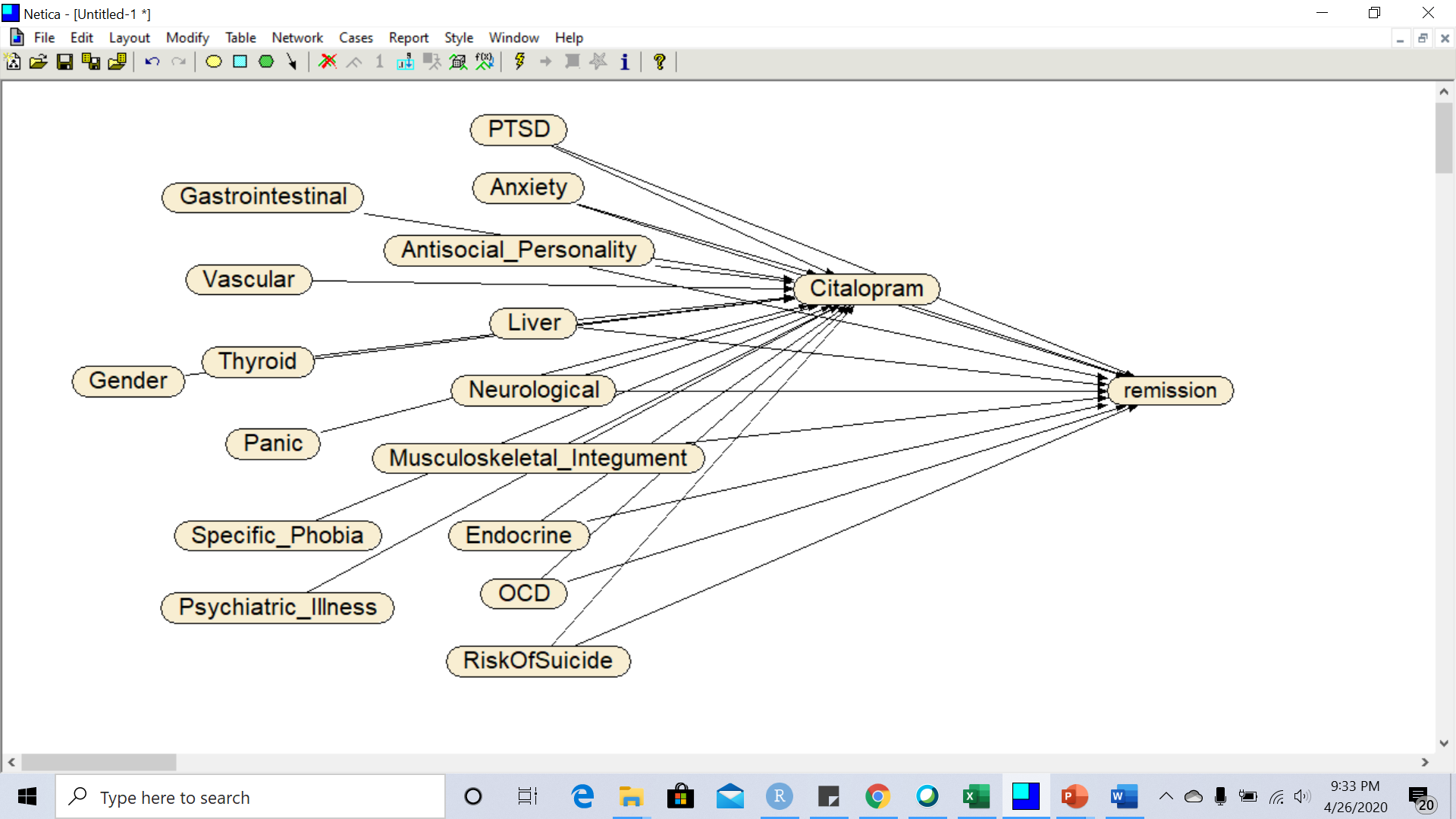
**Results:** In the LASSO regression of citalopram, the following independent variables were statistically significant: thyroid, PTDS, gender, neurological, OCD, risk of suicide, anxiety, psychiatric-Illness, vascular, musculoskeletal Integument, gastrointestinal, specific phobia, antisocial personality, panic, liver and endocrine disorders. Figure 2 shows the network model of variables that affect selection of citalopram.

**Figure 2: Variables that Affect Selection of Citalopram**



In the next LASSO regression, the variables that were statistically significant were risk of suicide, PTSD, anxiety, antisocial personality, liver, neurological, musculoskeletal integument, endocrine, citalopram, and OCD. In addition, citalopram was also a significant predictor of symptom remission. The combined LASSO regressions describe the network in Figure 3.

**Figure 3: Combined Network Model for Predicting Effect of Citalopram on Remission**



The net accuracy of predictions can be assessed from the percentage of variation explained by each of the regressions, as seen in Table 3:

**Table 3: Impact of Citalopram**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Independent Variables in the Model** | **Number of Covariates (Percent Reduced)** | **Impact of Citalopram on Remission** | **Percent of Variation Explained** | **10-fold Cross-Validated AROC** | **AROC divided by Degrees of Freedom** |
| Comprehensive: Gender, 32 Baseline Diagnoses, & Citalopram | 34 (100%) | 16(47.05%) | 0.532 | 61.00% | 1.79 |
| Only Parents of Citalopram & Citalopram | 16(100%) | 10(62.5%) | 0.514 | 57.00% | 3.56 |

**Limitations**

Like any statistical modeling effort, under powered analysis and analysis that is not cross-validated could lead to errors. We reduced the possibility of underpowered analysis by relying on large data set and cross-validating our findings. We also reduced the possibility of under-powering the analysis by focusing on comorbidities that occurred at least 100 times with stomach cancer. Many investigators need to analyze smaller data sets; in which it may not be possible to carry out the proposed analysis.

We contrasted two network-inspired methods of reducing number of variables stratified. We did not examine several other popular approaches for feature reduction. For example, we did not examine the performance of these network inspired methods against methods that identify the Markov Blanket of cancer or survival. This paper was not intended to compare various methods of feature reduction but to show the details of how Pearl’s backdoor method would work.

While we reduced the number variables stratified, we provide no guarantee that this is the minimum set of variables that should be stratified. If there are covariates that are not on the Markov blanket but mediate the effects of many other variables, then stratifying these covariates may make sense. To identify the minimum set of covariates that need to be stratified, it would help to derive the structure of the network model, a task that we wanted to avoid.

**DISCUSSION**

PT and PO reduced the number of variables stratified by 71% and 63%. These are large reductions in covariates stratified. Despite reduction in number of covariates stratified, the two methods had similar accuracy levels to the situation where all covariates were used. In predicting survival from stomach cancer, PT had 10-fold cross validated Area under the Receiver Operating Curve of 81.16%, versus 80.01% for PO and 80.14% for stratifying all covariates. These data confirm that PT and PO are more computationally efficient and roughly equally accurate as the comprehensive approach.

One implication of stratifying fewer variables is its impact on matching cases and controls. In high dimensional data, a large portion of cases are not matched to controls. In the literature, this is reported as “extent of overlap”. A number of investigators have examined methods of improving case/control matching including propensity scoring (Morgan and Harding 2006) and synthetic controls (Huberman and Langholz 1999; Vach and Blettner 1991). By reducing the number of variables stratified, the PT approach improved the percentage of cases matched from 25.3% in the comprehensive approach to 88.5% in the PT method. The higher percentage of cases matched increases the ability of generalizing from the cases in the study to patients with different combination of stomach cancer comorbidities.

The findings in this study suggest that PT method is practical and can be applied to high dimensional problems. The data suggests that reduction of stratification was accomplished with limited loss of accuracy. These findings need to be replicated in other settings and with other data before one has confidence in the data. If the findings hold true in other data, then PT is one of the several methods to enable the use of stratification in removing confounding in high-dimensional data.

**REFERENCES**

Abadie A, Imbens GW. Bias-corrected matching estimators for average treatment effects. J. Bus. Econ. Statist 2011, 29, 1–11.

Abadie A, Imbens GW. Large sample properties of matching estimators for average treatment effects. Econometrica 2006, 74, 235–267.

Alemi F, ElRafey A, Avramovic I. Covariate Balancing through Naturally Occurring Strata. Health Serv Res. 2016 Dec 14. doi: 10.1111/1475-6773

Alemi F, Zargoush M, Vang J. Using observed sequence to orient causal networks. Health Care Manag Sci. 2017 Dec; 20(4):590-599. doi: 10.1007/s10729-016-9373-3.

Aliferis CF, Tsamardinos I, Statnikov A. HITON: a novel Markov Blanket algorithm for optimal variable selection. AMIA Annu Symp Proc. 2003:21-5.

Atkinson AC. Optimum biased coin designs for sequential clinical trials with prognostic factors. Biometrika. 1982; 69(1): 61-7.

Benichou J. Methods of adjustment for estimating the attributable risk in case-control studies: a review. Stat Med. 1991 Nov; 10(11):1753-73.

Castillo E, Gutierrez JM, Hadi AS, Expert Systems and Probabilistic Network Models. Springer-Verlag, 1997.

Chen R, Sun J, Dittus RS, Fabbri D, Kirby J, Laffer CL, McNaughton CD, Malin B. Patient Stratification Using Electronic Health Records from a Chronic Disease Management Program. IEEE J Biomed Health Inform. 2016 Jan 4.

Confavreux C, Vukusic S. Age at disability milestones in multiple sclerosis. Brain. 2006 Mar;129(Pt 3):595-605.

de Graaf MA, Jager KJ, Zoccali C, Dekker FW. Matching, an appealing method to avoid confounding? Nephron Clin Pract. 2011; 118(4): c315-8.

de Vlaming R, Groenen PJ. The Current and Future Use of Ridge Regression for Prediction in Quantitative Genetics. Biomed Res Int. 2015; 2015: 143712

Frangakis CE, Rubin DB. Principal stratification in causal inference. Biometrics, 2002, 58, 21–29.

Fu, Shunkai, Desmarais MC. Markov blanket based feature selection: a review of past decade." In Proceedings of the world congress on engineering, vol. 1, pp. 321-328. 2010.

Fung R, Chang K. Weighing and integrating evidence for stochastic simulation in Bayesian networks,” in Uncertainty in Artificial Intelligence, M. Henrion, R. D. Shachter, L. N. Kanal, and J. F. Lemmer, Eds., 1990

Gyöngyösi M, Ploner M, Porenta G, Sperker W, Wexberg P, Strehblow C, Glogar D. Case-based distance measurements for the selection of controls in case-matched studies: application in coronary interventions. Artif Intell Med. 2002 Nov; 26(3):237-53.

Hansen PT. Full matching in an observational study of coaching for the SAT. J. Am. Statist. Ass. 2004, 99, 609–618.

Heckman JJ, Ichimura H, Todd P. Matching as an econometric evaluation estimator. Review of Economic Studies 1998, 65, 261-294.

Heritier S, Gebski V, Pillai A. Dynamic balancing randomization in controlled clinical trials. Stat Med. 2005; 24(24): 3729-41.

Hirano K, Imbens G, Ridder G. Efficient estimation of average treatment effects using the estimated propensity score. Econometrica 2003, 71, 1307–1338.

Ho DE, Imai K, King G, Stuart EA. Matching as nonparametric preprocessing for reducing model dependence in parametric causal inference. Polit. Anal. 2007, 15, 199–236.

Huang Z, Dong W, Duan H. A probabilistic topic model for clinical risk stratification from electronic health records. J Biomed Inform. 2015 Dec; 58:28-36.

Huberman M, Langholz B. Application of the missing-indicator method in matched case-control studies with incomplete data. Am J Epidemiol. 1999 Dec 15; 150(12):1340-5.

Imai K, Ratkovic M. Covariate balancing propensity score. J. R. Statist. Soc. B (2014), 76, Part 1, pp. 243–263.

Li G, Dai H, Tu Y. Identifying Markov Blankets Using LASSO Estimation

Li Z, D’Ambrosio B. Efficient inference in Bayes’ networks as a combinatorial optimization problem. International Journal of Approximate Inference, vol. 11, 1994.

Maathuis MH, ColoPO D. A generalized backdoor criterion. Annals of Statistics 2015, 43 1060-1088.

Marafino BJ, Boscardin WJ, Dudley RA. Efficient and sparse feature selection for biomedical text classification via the elastic net: Application to ICU risk stratification from nursing notes. J Biomed Inform. 2015 Apr; 54:114-20.

Mei K, Peng J, Gao L, Zheng NN, Fan J. Hierarchical classification of large-scale patient records for automatic treatment stratification. IEEE J Biomed Health Inform. 2015 Jul; 19(4):1234-45.

Moreno V, Martín ML, Bosch FX, de Sanjosé S, Torres F, Muñoz N. Combined analysis of matched and unmatched case-control studies: comparison of risk estimates from different studies. Am J Epidemiol. 1996 Feb 1; 143(3):293-300.

Morgan SL, Harding DJ. Matching Estimators of Causal Effects: Prospects and Pitfalls in Theory and Practice. Sociological Methods & Research 2006 35: 3

Opgen-Rhein R, Strimmer K. From correlation to causation networks: a simple approximate learning algorithm and its application to high-dimensional plant gene expression data. BMC Syst Biol. 2007 Aug 6; 1:37.

Pearl J. Causality: Models, Reasoning, and Inference. Cambridge University Press, 2000.

Pearl J. An introduction to causal inference. Int J Biostat. 2010 Feb 26; 6(2): Article 7.

Pearl J. Probabilistic reasoning in intelligent systems: networks of plausible inference. Morgan Kaufmann, 2014

Pirracchio R, Resche-Rigon M, Chevret S. Evaluation of the propensity score methods for estimating marginal odds ratios in case of small sample size. BMC Med Res Methodol. 2012 May 30;12:70

Pocock SJ, Simon R. Sequential treatment assignment with balancing for prognostic factors in the controlled clinical trial. Biometrics. 1975; 31(1):103-15.

Robins JM, Hernan MA, Brumback B. Marginal structural models and causal inference in epidemiology. Epidemiology 2000, 11, 550–560.

Robins JM, Rotnitzky A, Zhao LP. Analysis of semiparametric regression models for repeated outcomes in the presence of missing data. J. Am. Statist. Ass. 1995, 90, 106–121.

Rosenbaum PR, Rubin DB. Constructing a control group using multivariate matched sampling methods that incorporate the propensity score. Am. Statistn 1985, 39, 33–38.

Rosenbaum PR, Rubin DB. Reducing bias in observational studies using sub-classification on the propensity score. J. Am. Statist. Ass. 1984, 79, 516–524.

Rosenbaum PR, Rubin DB. The central role of the propensity score in observational studies for causal effects. Biometrika 1983, 70, 41–55.

Rosenbaum PR. A characterization of optimal designs for observational studies. J. R. Statist. Soc 1991. B, 53, 597–610.

Rosenbaum PR. Model-based direct adjustment. J. Am. Statist. Ass. 1987, 82, 387–394.

Rosenbaum PR. Optimal matching for observational studies. J. Am. Statist. Ass. 1989, 84, 1024–1032.

Rosenbaum PR. Sensitivity analysis for matched case-control studies. Biometrics. 1991 Mar; 47(1):87-100.

Rosenbaum PR. Sensitivity analysis for m-estimates, tests, and confidence intervals in matched observational studies. Biometrics. 2007 Jun; 63(2):456-64

Scott NW, McPherson GC, Ramsay CR, Campbell MK. The method of minimization for allocation to clinical trials: a review. Controlled Clinical Trials 2002; 23(6): 662-74.

Shachter RD. Probabilistic inference and influence diagrams. Operations Research, vol. 36, 1988.

Shen J, Lida L, Wong W. Markov Blanket Feature Selection for Support Vector Machines." In AAAI, vol. 8, pp. 696-701. 2008.

Shojaie A. and Michailidis G. (2010) Penalized Likelihood Methods for Estimation of Sparse High Dimensional Directed Acyclic Graphs, Biometrika, 97(3): 519-538.

Signorini DF, Leung O, Simes RJ, Beller E, Gebski VJ, Callaghan T. Dynamic balanced randomization for clinical trials. Stat Med. 1993; 12(24):2343-50.

Singh A, Nadkarni G, Gottesman O, Ellis SB, Bottinger EP, Guttag JV. Incorporating temporal EHR data in predictive models for risk stratification of renal function deterioration. J Biomed Inform. 2015 Feb; 53:220-8.

Smith JA, Petra T. Does Matching Overcome LaLonde’s Critique of Nonexperimental Estimators?’’ Journal of Econometrics 2005, 125:305-53.

Tan Y, Zhifa L. Feature selection and prediction with a Markov blanket structure learning algorithm." BMC bioinformatics 14, no. 17 (2013): A3.

McCullough J. Inference and Identification in Machine Learning and Application for Health Policy. Invited talk at George Mason University, Department of Health Administration and Policy, May 2, 2017.

Tran T, Luo W, Phung D, Harvey R, Berk M, Kennedy RL, Venkatesh S. Risk stratification using data from electronic medical records better predicts suicide risks than clinician assessments. BMC Psychiatry. 2014 Mar 14; 14:76.

Tripepi G, Jager KJ, Dekker FW, Zoccali C. Stratification for confounding--part 1: the Mantel-Haenszel formula. Nephron Clin Pract. 2010; 116(4):c317-21.

Vach W, Blettner M. Biased estimation of the odds ratio in case-control studies due to the use of ad hoc methods of correcting for missing values for confounding variables. Am J Epidemiol. 1991 Oct 15;134(8):895-907.

Wei LJ. Application of an Urn Model to Design of Sequential Controlled Clinical-Trials. J Am Stat Assoc. 1978; 73(363):559-63.

Yifeng Z, Luo J, Lin S. Classification using Markov blanket for feature selection. In Granular Computing, 2009, GRC'09. IEEE International Conference on, pp. 743-747. IEEE, 2009.

Yuan Y, Huang XL, Liu SY. A Bayesian response-adaptive covariate-balanced randomization design with application to a leukemia clinical trial. Stat Med. 2011; 30(11): 1218-29.

Zeng Y, Jian L, Lin S. Classification using Markov blanket for feature selection. In Granular Computing, 2009, GRC'09. IEEE International Conference on, pp. 743-747. IEEE, 2009.