**Using Regression to Learn Causal Networks: Application to Analysis of Cost Overruns in Bundled Payments**

Version of November 11, 2016

Farrokh Alemi, PhD1

Amr Elrafey1

Larry Tang PhD2

1George Mason University, Department of Health Administration and Policy

2George Mason University, Department of Statistics

**Abstract**

**Objective**: This paper shows how Poisson regression can be used for causal analysis of data in electronic health records. **Methods:** Data were simulated to describe causes of cost overruns within the Center for Medicare and Medicaid’s bundle payment for hip fracture. The paper examines if the network that generated the data could be recovered from the simulated data. First, a partial sequence among the variables was deduced from the definition of the variables, creating 5 timeframes: Time 1: Pre-hospital Discharge variables, Time 2: Hospital Discharge, Time 3: Post-acute Institutional variables, Time 4: Outpatient/Hospice variables, and Time 5: Bundle Payment. Stepwise Poisson regression was done, where at every step variables in the next time period were added to the analysis. At each step, this analysis identifies parents in the Markov Blanket (PMB) of a variable. We also verified if the statistically significant associations identified in the data were causal by conducting an additional, smaller, Poisson regression of each variable and its PMB. Interactions that were statistically significant and verified to be causal were included in subsequent Poisson regression. A directed acyclical graph was organized for all relationships that remained statistically significant at the last time period. **Results:** The initial stepwise Poisson regression identified x links that did not exist in the original network and missed 1 link that did exist; the subsequent test of causality removed all excess links. The recovered network differed from the original network in only one out 156 possible links. **Conclusions:** It is possible to use multivariate Poisson regression to construct an acyclical directed network. The procedure is especially suitable for discovery of cost overruns in bundled payments.**Introduction**

 There are different methods of learning causal networks from data. These include two broad approaches: (1) search and scoring and (2) conditional dependency algorithms. In the search and scoring method, one examines different alternative network structures until one is found that fits the data best [[[1]](#endnote-2)–[[2]](#endnote-3),[[3]](#endnote-4),[[4]](#endnote-5),[[5]](#endnote-6),[[6]](#endnote-7),[[7]](#endnote-8)]. These algorithms include, Taboo [[[8]](#endnote-9),[[9]](#endnote-10)], Sopleq [[[10]](#endnote-11)], EQ [[[11]](#endnote-12)], Maximum Spanning Tree [[[12]](#endnote-13)], and Taboo Order [[[13]](#endnote-14)]. The main weakness of search and score methods is that many possible alternative network structures are possible, and a finding the suitable one can be time consuming. The possible number of unique network structures with n variables grows exponentially as n increases [[[14]](#endnote-15)]; for n = 3, it is 25; for n = 5, it is 29,000; and for n = 10, it is approximately 4.2×1018 [[[15]](#endnote-16)]. Clearly, even search and scoring methods require procedures to rule out potentially thousands of network layouts. Some investigators have proposed limiting the number of causes for any effect in the network [[[16]](#endnote-17),[[17]](#endnote-18)]. Obviously, such arbitrary limitations, while conducive to computation efficiency, may lead to erroneous causal models. For these situations, conditional dependency algorithms may provide some advantages.

 In conditional dependency algorithms, the conditional independence among triplets of variables are used to determine the possible cause and effect relationships [[[18]](#endnote-19)]. Some examples of these algorithms are: Three Phase Dependency Analysis [[[19]](#endnote-20)]; the PC algorithm [[[20]](#endnote-21)]; the Poly-tree Recovery algorithm [[[21]](#endnote-22)]; the Inductive Causation algorithm [[[22]](#endnote-23)]; Sprites, Glymour, and the Scheines algorithm [[[23]](#endnote-24)]; the Grow-Shrink algorithm [[[24]](#endnote-25)]; and Fast Casual Inference algorithm [[[25]](#endnote-26)]. The principle behind these algorithms is to identify a triplet of variables with a common effect configuration. In a network that does not allow cycles, any triplet of variables X, Y and Z can relate to each other in one of three possible arrangements: causal chain, common cause, or common effect [[[26]](#endnote-27)]. When a common effect, sometimes called collider, configuration exists, we can infer the direction of the links in the network. A common effect configuration exists when two previously independent variables X and Y are conditionally dependent on Z. This switch from independence to conditional dependence is the core building block behind various constraint based algorithms. These algorithms repeatedly test for this switch across triplets of variables.

**Poisson Regression**

 Poisson regression is an alternative to existing algorithms for learning causal networks [[[27]](#endnote-28)]. Poisson regression assumes that the underlying process has a Poisson distribution in which the expected value and standard deviations are equal. The assumptions of Poisson process are:

1. The probability of more than one event occurring in a short period of time is negligible. Hospitalization for multiple diseases may violate this assumption unless we focus on primary reason for hospitalization.
2. The rate of occurrence of the events is constant. This assumption is violated among patients infected with a contagious disease as the rate of occurrence of the event increases in subsequent samples.
3. The probability of an event occurring in a small time interval does not depend on the length of time since the previous event. If the prior occurrence of an event affects its future occurrence, e.g. when hospitalization raises the probability of subsequent hospitalization, then this assumption is violated.

If **C**{\displaystyle \mathbf {x} \in \mathbb {R} ^{n}} is a vector of independent variables, T the treatment variable, then the Poisson Regression of outcome Y on the independent variables takes the form:

{\displaystyle \log(\operatorname {E} (Y\mid \mathbf {x} ))=\alpha +\mathbf {\beta } '\mathbf {x} ,}

In Poisson regression, associations among any pair of variables are identified through regressing log of count of combination of variables on main and pair-wise interactions of the variables. In this type of regression the association of two variables is measured after stratifying all other variables. This feature makes Poisson regression similar in concept to constraint based methods of learning network structure. Obviously, Poisson regression is focused on learning associations and not causal effects. All causal effects suggests existence of associations but not vice versa. So, some of the associations identified in Poisson regression must be ignored. In this paper we show how to accept some, but ignore other, associations found in Poisson regression.

**Sequence of Data**

In electronic health records, the sequence among the data can be easily discovered using several different methods. Sometimes, the definition of the variables tell us that one variable is measured before another. For example, demographic variables are measured before other covariates since they are features that the patient acquired on birth. For another example, outcomes are usually measured after other covariates or treatment and not before.

 Sometimes, the timing of data collection indicates the sequence of the variables. Thus, we may learn sequence among diagnoses of a patient by examining the timing of these events. Medical history, typically, are recorded over time revealing timing of events. Still other times, timing of events can be deduced through age at which these events typically occur. Thus, one may deduce that cardiovascular events usually occur in 60s and 70s while Alzheimer occurs in 80s and 90s. The age at which various events are most likely to occur may reveal the sequence among these events. Finally, there are also empirical tests of sequence. These include conditional independence test of colliders which can establish that two or more causes occur before their common effect. Other tests like Probabilistic Contrast model, and Goodman and Kruskal error reduction methods can also be used to establish sequence [[[28]](#endnote-29)].

**Drawing a Network Based on Poisson Regression**

 Agresti [[[29]](#endnote-30) ] showed how hierarchical Poisson regression can be used to identify the associations among the variables. First, he found the best fit to data by progressively removing pair-wise interaction terms and re-examined the model’s goodness of fit to the data. A network was drawn by creating a node for each variable. For every association that remained in the best fit model, whether statistically significant or not, Agresti drew an arc between the nodes.

 An alternative approach, one that we prefer, is to map to the network only statistically significant pair-wise associations that have causal implications. Poisson regression examines pair-wise associations when all other variables are stratified. Poisson regression identifies the Markov Blanket of each variable. The Markov Blanket is the parents (direct causes), children (effects) and co-parents of a variable. It is identified when a statistically significant relationship exist among two variables despite stratification of all other variables. The Markov Blanket of a variable can be used to organize a network structure, especially if sequence information is available.

**Problems in Learning Networks from Poisson Regression**

 There are at least two ways that Poisson regression may identify non-causal associations: (1) stratifying common effects, and (2) showing correlations due to common causes. A common effect is when two or more causes have the same effect, e.g. in the network X🡪Y🡨Z, X and Z have the common effect Y. If multiple causes have a common effect, then stratifying the common effect will create an association among the causes [[[30]](#endnote-31),[[31]](#endnote-32)]. In our example, stratifying Y will create a correlation among X and Z. Since Poisson regression detects a relationship between two variables by stratifying all other variables it is likely that it will also stratify a common effect and report a non-causal association. To help avoid stratifying common effects, we have designed a stepwise procedure for Poisson regression. Note that all causes occur in time period prior to the effects; then the non-causal association always occurs prior to the time of the common effect. This suggests that a time-based, stepwise, procedure could avoid discovery of non-causal association due to stratifying of common effects. Variables at time period t are examined first, statistically significant interaction terms are kept, and all other interactions are discarded; thus, preventing these interactions from becoming significant in models constructed in later time periods.

The second type of non-causal associations are due to common causes. A common cause refers to the situation where one variable affects two or more other variables. For example, in this network X🡨Y🡪Z, Y is the common cause of X and Z. A common cause is likely to lead to co-variation among its effects (X and Z are likely to be correlated). Poisson regression may indicate an association even though there is no causal link in the original network that generated the data. To prevent these non-causal associations, we have designed a test of causality using the Parents of the Markov Blanket of variables added in the last time period. Pearl’s work on “do” operation and blocking backdoors suggests the basic idea [[[32]](#endnote-33)]. Pearl shows that causal impact can be calculated if the backdoors starting from the effect to the cause are blocked. Stratifying all parents in the Markov Blanket of a variable blocks all backdoors that could go from the effect to the cause. In particular, a smaller Poisson regression is done with all variables in the Markov Blanket of the effect. If the association remains significant, then it is likely to be causal. If not, then the interaction term is ignored in the subsequent time periods and the association is not shown as an arc in the discovered causal network.

**Source of Data**

 To demonstrate how Poisson Regression can be used to construct a network model, we simulated data based on the network in Figure 1. This network shows the relationship among various sources of cost overrun and whether bundle-payment of the Center for Medicare and Medicaid (CMS) was above or below the total cost. At end of the year, CMS reduces payments made to a hospital for treatment of hip fracture (DRG 469 or 470) and subsequent 90 days post-discharge costs, if these costs were higher than average. The following costs are included:

1. Physicians' services, P
2. Inpatient hospital services, H
3. Long-term care hospital services, LTH
4. Inpatient rehabilitation facility services, RF
5. Skilled nursing facility services, SNF
6. Home health agency services, HHA
7. Hospital outpatient services, HO
8. Outpatient therapy services, OT
9. Clinical laboratory services, CL
10. Durable medical equipment, DME
11. Part B drugs, PBD
12. Hospice, HO
13. Bundled payments, BP

 The Centers for Medicare and Medicaid excludes from the calculation of bundled payments unrelated costs, e.g. if the person was hospitalized for cancer, the cost would be excluded. Despite efforts to avoid unrelated costs, the management of services to patients paid in bundled amounts is often difficult. Hospital administrators have to improve the efficiency and quality of their own operations as well care at other organization so that the hospital does not lose funds as a consequence of bundled payments. For a hospital administrator to do so, he needs to understand causes of cost overrun, which could be a function of either pricing of the service, volume of the service, or inappropriate quality of care in any of the services. The network in Figure 1 shows a possible situation where a hospital administrator might find himself, where cost overruns in one area have led to total cost exceeding bundled payments. In these circumstances, Center for Medicare and Medicaid would reduce reimbursement to the hospital creating a financial incentive to find the causes of cost overrun and improve reimbursement.

**Figure 1: Network Used to Simulate Total Cost Exceeding CMS’s Bundled Payment**Nodes show cost overrun for CL = Clinical Laboratory, DME = Durable Medical Equipment, H = Hospital, P = Physician, LTH = Long Term Hospital, RF = Rehab Facility, SNF = Skilled Nursing Facility, HHA = Home Health Agency, HO = Hospital Outpatient, OT = Outpatient Therapy, PBD = Part B Drugs, HOS = Hospice,

BP = Bundle Payment



The arcs in the network show causes of cost exceeding bundled payments. Thus cost overrun in one service is shown to affect the probability of cost overrun in total 90 days. The hospital administrator, of course, is not aware of these causal relationships and must discover these relationships from reports on individual patients. In essence, he has to discover the causal network from the cases. We simulated 10,000 cases from network in Figure 1 using the Netica (version #) software from Norsys. In order to reduce errors due to Power of the test to detect the effect we made sure that each cause either more than doubled or reduced by more than ½ the probability of the effect. The code used to generate the network is available through the second author (AE). The data used in this analysis are provided online [[[33]](#endnote-34)].

**Determining the Sequence**

In this example, the determination of sequence is relatively easy. We know that the episodes starts with a hospital admission, during which the physician bill for their service, clinical laboratory tests are ordered and durable medical devices are placed in patient’s hip. So, DME, CL and P are the events that occur in the first time period. Hospital discharge, H, occurs in the second time period. Patients who are discharged from the hospital go to a series of post-acute services. These post-acute services are assumed to occur in time period 3. Discharge to these post-acute services will be primarily to outpatient services. We have assumed that no re-hospitalization occur and there are no cycles in the network. In addition, the way we simulated the data assumes that all causes occur prior to effects and not simultaneously with the effect. The final event is bundle payment which is assumed to occur 90 days after initial hospitalization.

**Stepwise Pass through Data**

 The assumptions of Poisson regression are more likely to be met if we examine counts of events, in this case count of combination of various cost overruns within the data. To accomplish this task we used the function ddply() from the package Plyr (version #) as follows:

|  |
| --- |
| *# Count* combination of variablesDatacounts = ddply(data,(1:13),nrow) |

A total of 2,660 combinations of the 13 variables in the simulation were identified. Among these, 1,133 combinations occurred only once in the database. Many (67%) of the theoretically possible combinations of 13 binary variables never occurred. Table 1 provides the top 20 combinations with highest counts. These data suggest that a Poisson process could have generated the count of combinations as the probability of any one of the combinations repeating is exceedingly small, less than 0.01.

**Table 1: 10 Combination of Cost Overruns in Different Services**Cost overruns for CL = Clinical Laboratory, DME = Durable Medical Equipment, H = Hospital, P = Physician, LTH = Long Term Hospital, RF = Rehab Facility, SNF = Skilled Nursing Facility, HHA = Home Health Agency, HO = Hospital Outpatient, OT = Outpatient Therapy, PBD = Part B Drugs, HOS = Hospice,

BP = Bundle Payment

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **CL** | **DME** | **H** | **P** | **LTH** | **RF** | **SNF** | **HHA** | **HO** | **OT** | **PBD** | **HOS** | **BP** | **Count** |
| 1 | 1 | 1 | 0 | 0 | 0 | 0 | 1 | 0 | 1 | 0 | 0 | 1 | 125 |
| 1 | 1 | 1 | 1 | 0 | 1 | 0 | 1 | 0 | 0 | 0 | 0 | 1 | 115 |
| 1 | 1 | 1 | 1 | 0 | 1 | 0 | 0 | 0 | 1 | 0 | 0 | 1 | 106 |
| 1 | 1 | 1 | 1 | 0 | 1 | 1 | 0 | 0 | 1 | 0 | 0 | 0 | 92 |
| 1 | 1 | 1 | 0 | 0 | 0 | 0 | 1 | 0 | 1 | 0 | 0 | 0 | 78 |
| 1 | 0 | 1 | 0 | 0 | 0 | 0 | 1 | 0 | 1 | 0 | 0 | 1 | 58 |
| 1 | 0 | 1 | 1 | 0 | 1 | 0 | 0 | 0 | 1 | 0 | 0 | 1 | 57 |
| 0 | 1 | 1 | 0 | 0 | 0 | 0 | 1 | 0 | 1 | 0 | 0 | 1 | 56 |
| 1 | 0 | 1 | 1 | 0 | 1 | 0 | 1 | 0 | 0 | 0 | 0 | 1 | 53 |
| 1 | 1 | 1 | 0 | 0 | 0 | 1 | 1 | 0 | 1 | 0 | 0 | 1 | 50 |

The Poisson regression is done through a time-based stepwise process, where variables are entered into the analysis in order of their occurrence. Details and related R code are provided in the Appendix. For example, The R code for analysis of variables at Time 1 is given by:

|  |
| --- |
| Insert |

Table 2 provides a summary of the results of Poisson regressions. In this Table, we report only findings of significance for pairwise interactions, main effects were part of the model but of no relevance to our analysis and are not reported here. If an interaction term was not included in the regression model, then the cell is shaded. Some interaction terms that show in initial analysis but not in subsequent analysis are eliminated from all subsequent analysis and therefore will have shaded cells in subsequent analysis. Interaction terms that were excluded because of timing also have shaded cells.

In Table 2, T1 reports the first Poisson regression. This regression includes the interactions among the 3 variables in Time 1: DME, P and CL. Note that none of the interactions among these 3 variables were significant. These interaction terms are excluded from subsequent analysis. At time 2, we add in the Hospital, H, variable. The model in Time 2 includes all interaction terms that were significant at Time 1 (none) and every possible interaction with variables in time 2. In Time 2 several statistically significant relationships are found. Note that H is a common effect of P, DME and CL, and if we had not done the stepwise process, the Poisson regression would have also found new relationships among the variables in Time 1.

**Table 2: Statistically Significant Relationships in Poisson Regression**

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Interaction Terms in Poisson Regression** | **T1** | **T2** | **H** | **T3** | **HHA**  | **SNF** | **RF**  | **PBD**  | **LTH** | **T4** | **OT** | **HO** | **HOS** | **T5** | **BP**  | **Recovered** | **Original** |
| P:CL | N |   |   |   |   |   |   |   |   |   |   |   |   |   |   | N | N |
| DME:P | N |   |   |   |   |   |   |   |   |   |   |   |   |   |   | N | N |
| DME:CL | N |   |   |   |   |   |   |   |   |   |   |   |   |   |   | N | N |
| P:H |   | Y | Y | Y |   |   |   |   |   | Y |   |   |   | Y |   | Y | Y |
| CL:H |   | Y | Y | Y |   |   |   |   |   | Y |   |   |   | Y |   | Y | Y |
| DME:H |   | Y | Y | Y |   |   |   |   |   | Y |   |   |   | Y |   | Y | Y |
| Other:HHA |   |   |   | Y | N  |   |   |   |   |   |   |   |   |   |   | N | N |
| P:HHA |   |   |   | Y | Y |   |   |   |   | Y  |   |   |   | Y  |   | Y  | Y  |
| Other:SNF |   |   |   | Y  |   | Y  | N  |   |   |   |   |   |   |   |   | N  | N  |
| P:SNF |   |   |   | Y  |   | Y  | Y  |   |   | Y  |   |   |   | Y  |   | Y  | Y  |
| H:SNF |   |   |   | Y  |   | Y  | Y  |   |   | Y  |   |   |   | Y  |   | Y  | Y  |
| Other:RF |   |   |   | Y  |   |   | N  |   |   | N  |   |   |   | N  |   | N  | N |
| P:RF |   |   |   | Y  |   |   | Y  |   |   | Y  |   |   |   | Y  |   | Y  | Y  |
| H:RF |   |   |   | Y  |   |   | Y  |   |   | Y  |   |   |   | Y  |   | Y  | Y  |
| Other:PBD |   |   |   | Y  |   |   |   | N |   | N |   |   |   | N |   | N | N |
| Other:LTH |   |   |   | Y  |   |   |   |   | N  | N  |   |   |   | N  |   | N  | N |
| P:LTH |   |   |   | Y  |   |   |   |   | Y  | Y  |   |   |   | Y  |   | Y  | Y  |
| H:LTH |   |   |   | Y  |   |   |   |   | Y  | Y  |   |   |   | Y  |   | Y  | Y  |
| Other:OT |   |   |   |   |   |   |   |   |   | Y  | N  |   |   |   |   | N  | N |
| HHA:OT |   |   |   |   |   |   |   |   |   | Y  | Y  |   |   | Y  |   | Y  | Y  |
| SNF:OT |   |   |   |   |   |   |   |   |   | Y  | Y  |   |   | Y  |   | Y  | Y  |
| RF:OT |   |   |   |   |   |   |   |   |   | Y  | Y  |   |   | Y  |   | Y  | Y  |
| PBD:OT |   |   |   |   |   |   |   |   |   | Y  | Y  |   |   | Y  |   | Y  | Y  |
| LTH:OT |   |   |   |   |   |   |   |   |   | Y  | Y  |   |   | Y  |   | Y  | Y  |
| Other:HO |   |   |   |   |   |   |   |   |   | Y  |   | N  |   | N  |   | N  | N |
| PBD:HO |   |   |   |   |   |   |   |   |   | Y  |   | Y  |   | Y  |   | Y  | Y  |
| LTH:HO |   |   |   |   |   |   |   |   |   | Y  |   |   | N  |   |   | N  | N  |
| Other:HOS |   |   |   |   |   |   |   |   |   | Y  |   |   | N  |   |   | N | N |
| LTH:HOS |   |   |   |   |   |   |   |   |   | Y  |   |   | Y  | Y  |   | Y  | Y  |
| H:HOS |   |   |   |   |   |   |   |   |   | Y  |   |   | Y  | Y  |   | Y  | Y  |
| PBD:BP |   |   |   |   |   |   |   |   |   |   |   |   |   | N  | N  | N  | Y  |
| SNF:BP |   |   |   |   |   |   |   |   |   |   |   |   |   | Y  | Y  | Y  | Y  |
| HHA:BP |   |   |   |   |   |   |   |   |   |   |   |   |   | Y  | Y  | Y  | Y  |
| OT:BP |   |   |   |   |   |   |   |   |   |   |   |   |   | Y  | Y  | Y  | Y  |
| HO:BP |   |   |   |   |   |   |   |   |   |   |   |   |   | Y  | Y  | Y  | Y  |
| HOS:BP |   |   |   |   |   |   |   |   |   |   |   |   |   | Y  | Y  | Y  | Y  |
| H:BP |   |   |   |   |   |   |   |   |   |   |   |   |   | Y  | Y  | Y  | Y  |
| Other:BP |   |   |   |   |   |   |   |   |   |   |   |   |   | Y  | N  | N | N |
| Main effects are not listed in this Table |

In Time 1 and 2, the relationships that have a statistically significant relationship with H, are in the Markov Blanket of H. These relationships can be tested by stratifying all other statistically significant relations and examining if the effect remains statistically significant. This is done by conducting a smaller regression involving H and its current Markov Blanket. In this case, this is the same regression that was done for Time 2.

Table 2 shows what happens when we move to Time 3. We include in the model all variables that remained statistically significant at Time 2. In addition, a set of 5 new variables are introduced in Time 3. The interaction among these 5 variables and all other variables are included in the Poisson regression model. The R code for this regression is given below, where we have highlighted the new Time 3 related interaction terms:

|  |
| --- |
| u=glm(V1 ~ DME + H + P + PBD + SNF + HHA + RF + LTH + DME:H + CL:H + P:H + CL:SNF + DME:SNF + H:SNF + P:SNF + CL:PBD + DME:PBD + H:PBD + P:PBD + CL:RF + DME:RF + H:RF + P:RF + CL:LTH + DME:LTH + P:LTH + H:LTH + CL:HHA + DME:HHA + P:HHA + H:HHA , data=t, family=poisson) |

The result of this regression is shown in the Appendix and also summarized in Table 2. Note that the analysis has identified a number of new associations that did not exist in the original network, which was used to simulate the data. Figure 2 shows these errors in red. At this point, we do smaller Poisson regressions involving variables in the Markov Blanket of variables introduced in Time 3. Many of the erroneous relationships are removed from the data by these smaller Poisson regressions.

In the next forward step, we include variables that remained significant in Time 3 and add in interactions with variables in Time 4. Note again that the analysis of Time 4 identifies a number of associations that exist in the data but not in the network that simulated the data. In Figure 2 these associations are shown in red.

**Figure 2: Errors that Remain in Different Time periods**Black Lines show Correct Links, Red Lines Incorrect Links

|  |  |
| --- | --- |
|  |  |
|  |  |
|  |  |

 In the last forward step, we include Bundled Payment, BP, in the analysis. The analysis includes all pair-wise interactions that were both statistically significant and large in the previous analysis. In addition, it includes all pair-wise interactions among BP and all previous variables. Table 2 shows the summary of findings and Appendix provides the details. Again, Poisson regression of the parents of Markov Blanket of the variables introduced in this time period removes the errors.

**Validity**

The analysis has nearly identified the network that was used to generate the data. To test the accuracy of the analysis we examined the frequency of discrepancies between discovered and original network. A directed arc could be absent in the original and present in the discovered network, and vice versa. Table 7 summarizes the frequency with which arcs were absent and present within the two networks.

**Table 7: Correspondence between Discovered and Original Networks**

|  |  |  |
| --- | --- | --- |
|  | **Present in Discovered Network** | **Absent in Discovered Network** |
| **Present in Original Network** | xx | xx |
| **Absent in Original Network** | xx | xx |

These data can also be reported using sensitivity and specificity of the discovered network. These data suggest that the discovered network was sensitive (xx), meaning that it identified most of the directed arcs in the original network accurately. It was also specific (xx), meaning that the arcs it identified also existed in the original network.

**Discussion**

 This paper has demonstrated that tools used for multivariate data analysis can be used to detect network structure and parameters. We have shown how forward and backward passes through the data reduce the errors in associations discovered through Poisson regression. In the forward steps, we showed how associations that exist because of collider effects are prevented to enter the model. In the backward steps, we showed that progressive test of causal impact improves the accuracy of Poisson regression. In this example, the discovered model had high sensitivity and specificity encouraging the use the proposed forward and backward passes.

 The procedure used here are similar to procedures proposed by Shojaie and colleagues [[[34]](#endnote-35)]. Like them, we used partial sequence among the variables to discover network structures and parameters. Unlike their approach, we relied on Poisson regression and not logistic or linear regression. Furthermore, we devised a forward and backward pass that removes most of the non-causal associations; i.e. associations that exist in the data but not in the causal model that generated the data. These include associations that exist because of stratifying a common effect or because of co-variation of several effects due to a common cause.

 There are a number of limitations in the procedures used in this paper. This paper has not proven that the approach will always be accurate. The accuracy has only been shown in one example. We have not established that Poisson regression will be accurate in every type of network. Additional simulation studies are needed so that one can anticipate when Poisson regression will be effective and why. The paper generated a large sample size and large effects for impact of each cause. These conditions may not always be present and will distort the findings of Poisson regression.

**References**

1. Heckerman D, Geiger D, Chickering D. Learning Bayesian networks: the combination of knowledge and statistical data. Mach Learn 1995, 20(3):197–234. [↑](#endnote-ref-2)
2. Cooper G, Herskovits E. A Bayesian method for the induction of probabilistic networks from data. Mach Learn 1992; 7:299–347. [↑](#endnote-ref-3)
3. Buntine W. Theory refinement on Bayesian networks. In: D’Ambrosio B, Smets P, Bonissone P (eds) 7th conference uncertainty in artificial intelligence, 1991, vol 26. Morgan Kaufmann, Los Angeles, pp 52–60 [↑](#endnote-ref-4)
4. Lam W, Bacchus F. Theory refinement on Bayesian networks. Comput Intell 1994, 10:269–293 [↑](#endnote-ref-5)
5. Wallace C, Korb K, Dai H. Causal discovery via MML. International Conference on Machine Learning, 1996, pp. 516–524 [↑](#endnote-ref-6)
6. Akaike H. A new look at the statistical model identification. IEEE Trans Autom Contr 1974, 19(6):716–723 [↑](#endnote-ref-7)
7. Schwarz G. Estimating the dimension of a model. Annals Stat, 1978: 6 [↑](#endnote-ref-8)
8. Chickering D, Geiger D, Heckerman D. Learning Bayesian networks: Search methods and experimental results. In Fifth International Workshop on Artificial Intelligence and Statistics, 1995, 112– 128. [↑](#endnote-ref-9)
9. Friedman N, Nachman I, Pe’er D. Learning Bayesian network structure from massive datasets: The “sparse candidate” algorithm. In Fifteenth Conference on Uncertainty in Artificial Intelligence (UAI- 99), 1999. [↑](#endnote-ref-10)
10. Naïm P, Wuillemin PH, Leray P, Pourret O, Becker A. Réseaux Bayésiens. Eyrolles, 2011. [↑](#endnote-ref-11)
11. Munteanu P, Bendou M. The EQ framework for learning equivalence classes of Bayesian networks. In Proceedings of the 2001 IEEE International Conference on Data Mining. 2001, 417–424. [↑](#endnote-ref-12)
12. Lam W, Bacchus F. Learning Bayesian belief networks: An approach based on the MDL principle. Computational intelligence, 1994; 10(3), 269–293. [↑](#endnote-ref-13)
13. Jouffe L, Munteanu P. New Search Strategies for Learning Bayesian Networks. In: June 2001. Compiègne (France). [↑](#endnote-ref-14)
14. Robinson W. Counting unlabeled acyclic digraphs, ser. Lecture notes in mathematics. Springer-Verlag, 1977, no. 622. [↑](#endnote-ref-15)
15. Cooper GF, Herskovits E. A Bayesian method for the induction of probabilistic networks from data. Machine Learning, vol. 9, pp. 309–347, 1992. [↑](#endnote-ref-16)
16. Friedman N; Linial M, Nachman I, Pe'er D. Using Bayesian Networks to Analyze Expression Data. Journal of Computational Biology 2000, 7 (3-4): 601–620. [↑](#endnote-ref-17)
17. Friedman N; Geiger D, Goldszmidt M. Bayesian Network Classifiers". Machine Learning 1997: 29 (2-3): 131–163. [↑](#endnote-ref-18)
18. Korb KB, Nicholson AE. Bayesian artificial intelligence. Chapman and Hall, 2003. [↑](#endnote-ref-19)
19. Cheng J, Greiner R, Kelly J, Bell DA, Liu E. “Learning Bayesian networks from data: an information-theory based approach,” The Artificial Intelligence Journal, vol. 137, pp. 43–90, 2002. [↑](#endnote-ref-20)
20. Spirtes P, Glymour C, Scheines R. Causation, Prediction, and Search. Springer-Verlag, 2000 [↑](#endnote-ref-21)
21. Pearl J. Probabilistic Reasoning in Intelligent Systems. Morgan Kaufmann, 1988. [↑](#endnote-ref-22)
22. Verma T, Pearl J. Equivalence and synthesis of causal models. UCLA, Tech. Rep., 1990. [↑](#endnote-ref-23)
23. Spirtes P, Glymour C, Scheines R. Causation, Prediction, and Search. Springer-Verlag, 2000 [↑](#endnote-ref-24)
24. Margaritis D. Learning Bayesian network model structure from data. Ph.D. dissertation, Carnegie Mellon University, 2003. [↑](#endnote-ref-25)
25. Spirtes P. An anytime algorithm for causal inference,” Carnegie Mellon University, Tech. Rep., 2000 [↑](#endnote-ref-26)
26. Jensen FV. Introduction to Bayesian networks. Springer-Verlag Telos, 1996. [↑](#endnote-ref-27)
27. Preston DL. Poisson regression in epidemiology. In: Armitage P, Colton T, editor. Encyclopedia of Biostatistics. 2. Vol. 6. Chichester: John Wiley & Sons; 2005. pp. 4124–4127. [↑](#endnote-ref-28)
28. Alemi F, Zargoush M, Vang J. Using observed sequence to orient causal networks. Health Care Manag Sci. 2016 Jul 30. [↑](#endnote-ref-29)
29. Agresti, A. Categorical Data Analysis. Wiley-Interscience, 2002. [↑](#endnote-ref-30)
30. Pearl J. Probabilistic reasoning in intelligent systems: networks of plausible inference. Morgan Kaufmann 1988. [↑](#endnote-ref-31)
31. Pearl J. Fusion, Propagation and Structuring in Belief Networks. Artificial Intelligence 1986. 29 (3): 241–288. [↑](#endnote-ref-32)
32. Pearl J. An introduction to causal inference. Int J Biostat. 2010 Feb 26;6(2):Article 7. [↑](#endnote-ref-33)
33. See <http://openonlinecourses.com/causalanalysis/simulated%20bundled%20data.csv> accessed on October 17, 2016 [↑](#endnote-ref-34)
34. Tan KM, Witten D, Shojaie A. The cluster graphical lasso for improved estimation of Gaussian graphical models. Comput Stat Data Anal. 2015 May;85:23-36. [↑](#endnote-ref-35)