# A Study in Causal Discovery from Population-Based Infant Birth and Death Records

**Subramani Mani, MBBS, MS,** and **Gregory F. Cooper, MD, PhD**
{mani,gfc}@cbmi.upmc.edu
Center for Biomedical Informatics, Suite 8084, Forbes Tower, Meyran Avenue,
University of Pittsburgh, Pittsburgh PA 15213

*In the domain of medicine, identification of the causal factors of diseases and outcomes, helps us formulate better management, prevention and control strategies for the improvement of health care. With the goal of exploring, evaluating and refining techniques to learn causal relationships from observational data, such as data routinely collected in healthcare settings, we focused on investigating factors that may contribute causally to infant mortality in the United States. We used the U.S. Linked Birth/Infant Death dataset for 1991 with more than four million records and about 200 variables for each record. Our sample consisted of 41,155 records randomly selected from the whole dataset. Each record had maternal, paternal and child factors and the outcome at the end of the first year—whether the infant survived or not. For causal discovery we used a modified Local Causal Discovery (LCD2) algorithm, which uses the framework of causal Bayesian Networks to represent causal relationships among model variables. LCD2 takes as input a dataset and outputs causes of the form* variable $X$ causes variable $Y$. *Using the infant birth and death dataset as input, LCD2 output nine purported causal relationships. Eight out of the nine relationships seem plausible. Even though we have not yet discovered a clinically novel causal link, we plan to look for novel causal pathways using the full sample after refining the algorithm and developing a more efficient implementation.*

## INTRODUCTION

The most useful explanation of a phenomenon is often a description of the underlying causal processes [1]. This is particularly true in the domain of medicine where identification of the causal factors of a disease can influence treatment planning, as well as the development of intervention strategies for disease prevention and control.

Well designed experimental studies, such as randomized control trials, are typically employed in ascertaining causal relationships. Here the value of the variable postulated to be *causal* is set randomly and its effects measured. These studies are appropriate in certain situations, for example, animal studies and studies involving human subjects that have undergone a thorough procedural and ethical review. Experimental studies may not, however, be feasible in many contexts due to ethical, logistical, or cost considerations. These practical limitations of experimental studies heighten the importance of exploring, evaluating and refining techniques to learn causal relationships from observational data, such as data routinely collected in healthcare settings. The goal is not to replace experimental studies, which are extremely valuable in science, but rather to augment and guide experimental studies when feasible.

The study reported here focused on investigating factors that may contribute causally to infant mortality in the United States. Infant mortality is one of the most important public health problems in the U.S. [2]. International comparisons based on data from the United Nations statistical office for the year 1991 show that there are 21 countries in the world with lower infant mortality rates than the United States. Japan had the lowest rate of 4.4, while the US rate was 8.9 [3].

This paper introduces an algorithm called LCD2 for efficiently searching for possible causal relationships that are suggested by large observational databases. Results are reported of applying LCD2 to an infant birth and death dataset.

## METHODS

### Infant Birth and Death Dataset

We used the U.S. Linked Birth/Infant Death dataset for 1991 [4]. This dataset consists of information on all the live births in the United States for the year 1991. It also has linked data for infants who died within one year of birth. More than two hundred variables containing various maternal, paternal, fetal and infant parameters are available. For the infants who died within the first year, additional data on mortality, including cause of death, is reported. The records total more than four million and the infant death record number is 35,496. We selected a random subset of 41,155 cases for use in the current study. We did so in order to limit the computational time complexity of searching for causal patterns in the data. A total of 87 variables were selected after eliminating redundant variables and variables not of clinical interest, such as ID number. Table 1 provides support that our sample is representative of the whole infant birth and death dataset for the year 1991.

### Assumptions for Causal Discovery

In the research reported here, we use causal Bayesian networks to represent causal relationships among model variables. This section provides a brief introduction to causal Bayesian networks, as well as a description of the assumptions we used to apply these networks for causal discovery.

1091—8280/99/$5.00 

Table 1: Sample of this study compared with the whole infant birth and death linked dataset for the year 1991*

| Attribute | State | Population (n = 4,146,555) n | % | Sample (n = 41,155) n | % |
|---|---|---|---|---|---|
| Infant Outcome | Lived | 4,111,059 | 99.14 | 40,818 | 99.18 |
| | Died | 35,496 | 0.86 | 337 | 0.82 |
| Child Gender | Male | 2,121,836 | 51.17 | 21,001 | 51.03 |
| | Fem. | 2,024,719 | 48.83 | 20,154 | 48.97 |
| Race of Mother | White | 3,264,230 | 78.72 | 32,480 | 78.92 |
| | Black | 693,990 | 16.74 | 6788 | 16.49 |
| | Other | 188,335 | 4.54 | 1887 | 4.59 |

*The 95% confidence interval on the difference in the population and sample proportions is (-0.006, 0.006) or tighter, suggesting that the sample is representative of the population.

A causal Bayesian network (or *causal network* for short) is a Bayesian network in which each arc is interpreted as a direct causal influence between a parent node (variable) and a child node, relative to the other nodes in the network [5]. Figure 1 illustrates the structure of a hypothetical causal Bayesian network structure, which contains five nodes. Due to limited space, the states of the nodes and the probabilities that are associated with this structure are not shown. The causal network structure in Figure 1 indicates, for

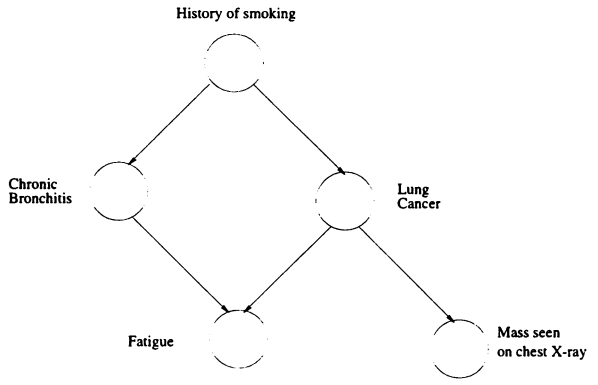


Figure 1: A hypothetical causal Bayesian network structure

example, that a *history of smoking* can causally influence whether *lung cancer* is present, which in turn can causally influence whether a patient experiences *fatigue* or presents with a *visible mass on chest X-ray*. The **causal Markov condition** gives the independence relationships[1] that are specified by a causal Bayesian network:

> A variable is independent of its non-descendants (i.e., non-effects) given just its parents (i.e., its direct causes).

According to the Markov condition, the causal network in Figure 1 is representing that the chance of a *mass seen on chest X-ray* will be independent of a *history of smoking*, given that we know whether *lung cancer* is present or not. While the causal Markov condition specifies independence relationships among variables, the **causal faithfulness condition** specifies *dependence* relationships:

> Variables are independent only if their independence is implied by the causal Markov condition.

For the causal network structure in Figure 1, three examples of the causal faithfulness condition are (1) *history of smoking* and *lung cancer* are probabilistically dependent, (2) *history of smoking* and *mass seen on chest X-ray* are dependent, and (3) *mass seen on chest X-ray* and *fatigue* are dependent. The intuition behind that last example is as follows: a *mass seen on chest X-ray* increases the chance of *lung cancer* which in turn increases the chance of *fatigue*; thus, the variables *mass seen on chest X-ray* and *fatigue* are expected to be probabilistically dependent. In other words, the two variables are dependent because of a common cause (i.e., a confounder). The causal Markov and faithfulness conditions describe *probabilistic* independence and dependence relationships, respectively, that are represented by a causal Bayesian network. In causal discovery, we do not know the probabilistic relationships among variables precisely, because we only have a finite amount of data. Thus, we make the following **statistical testing assumption**:

> A statistical test performed to determine independence (or alternatively dependence) given a finite dataset on population P will provide the same answer as when it is applied using an infinite dataset on P.

The greater the number of records in a dataset, the more likely it is that the statistical testing assumption will hold. Fortunately, the infant birth and death dataset contains a large number of records.

### An Algorithm for Causal Discovery

In this section, we introduce a causal discovery algorithm called LCD2[2]. LCD2 assumes the following:

**Assumption 1:** The causal Markov condition
**Assumption 2:** The causal faithfulness condition
**Assumption 3:** The statistical testing assumption
**Assumption 4:** There is a variable $W$ (called the instrumental variable) that is not caused by any other measured variable in the dataset.

[1] We use the terms *independence* and *dependence* in this section in the standard probabilistic sense.

[2] The algorithm is called LCD2, because it is an extension of the LCD algorithm that is described in [6]. The LCD algorithm only performed tests 1, 2, 3, and 6 that are described in this section

Before introducing the LCD2 algorithm, we define some terms. Let Independent$_T(A, B)$ denote that $A$ and $B$ are independent according to test $T$ applied to our dataset. Let Independent$_T(A, B$ given $C)$ denote that $A$ and $B$ are independent given $C$, according to $T$. Finally, let Dependent$_T(A, B)$ denote that $A$ and $B$ are dependent according to $T$.[3] Suppose that the following causal relationships are a valid model of nature: Given Assumptions 1–4 above, which we will

$$W \rightarrow X \rightarrow Y$$

Figure 2: A causal model in which $W$ causes $X$, and $X$ causes $Y$

assume in the remainder of this section, the following independence and dependence test results will hold:

Test$_1$. Dependent$_T(W, X)$
Test$_2$. Dependent$_T(X, Y)$
Test$_3$. Dependent$_T(W, Y)$
Test$_4$. Dependent$_T(W, X$ given $Y)$
Test$_5$. Dependent$_T(X, Y$ given $W)$
Test$_6$. Independent$_T(W, Y$ given $X)$

As proven in [6], Test$_1$ through Test$_6$ also will hold if there is a hidden variable causally influencing (i.e., confounding) $W$ and $X$, possibly in conjunction with $W$ causally influencing $X$ directly (see Figure 3).

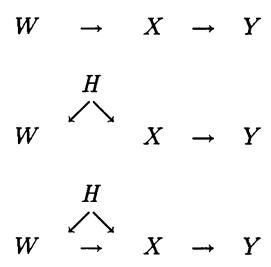


Figure 3: Causal models in which $X$ causes $Y$, and $W$ and $X$ are dependent due to some combination of (1) $W$ causing $X$ and (2) $W$ and $X$ being confounded by a hidden variable (or variables) represented by $H$.

Thus, as illustrated in Figure 3, if $X$ is causally influencing $Y$, and $X$ and $Y$ are not confounded, then Test$_1$ through Test$_6$ will hold. Conversely, as also proven in [6], if $X$ and $Y$ are both being causally influenced by one or more hidden variables (i.e., being confounded), then it follows from Assumptions 1–4 that Test$_6$ will not hold. Therefore, under Assumptions 1–4, if Test$_1$ through Test$_6$ hold, then (1) $X$ is causally influencing $Y$, and (2) $X$ and $Y$ are not confounded by a hidden variable.

[3] Although the three tests in this paragraph should technically be distinguished from each other by using separate labels, such as *T1*, *T2*, and *T3*, for simplicity of notation we use a single label $T$.

The LCD2 algorithm applies Test$_1$ through Test$_6$ in exploring a database for possible causal relationships. The algorithm is given as input a variable $W$; more generally we could call the algorithm with $r$ number of such $W$ variables. Given a variable $W$, the algorithm performs Test$_1$ through Test$_6$ for each pair of measured variables $X$ and $Y$ in the database. It uses simple variations of the Independence and Dependence tests described in [6]; both tests are $O(m)$ time complexity, where $m$ is the number of records (cases) in the database. If all six tests are passed, LCD2 outputs that $X$ causally influences $Y$ (under Assumptions 1–4), and it displays the probability distribution of $Y$ given $X$.

Traditional statistical approaches using $\chi^2$ tests or logistic regression can establish dependence between variables. Likewise, machine learning algorithms such as decision tree learners (e.g., C4.5 and CART), rule inducers (e.g., C4.5Rules and FOCL) and neural networks can build useful domain models from data and capture the inter-dependence among the variables. But none of these techniques is intended to establish causal relationships of the form $X$ causally influences $Y$.

The formalism of *structural equation models* (SEMs) [7], attempts to estabish causality, going beyond correlation and dependence. The emphasis in SEM research is on hypothesis testing of manually specified models, rather than on automated search over the space of models. Typically the SEM assumes linear relationships (with statistical noise) among the model variables; modeling with discrete variables is problematic. A discussion of the philosophical literature on causality is beyond the scope of this paper. For a detailed discussion of the relationship between statistical association and causation, including philosophical issues see for example [8] and [1].

Earlier research on learning Bayesian networks from data ([9], [10]) has simultaneously modeled all the causal relationships among the model variables. The LCD2 algorithm searches only for pairwise causal relationships. Thus, LCD2 trades off completeness for efficiency. In particular, if there are $n$ variables in the database, the time complexity of LCD2 is $O(mn^2r)$, where $m$ is the number of records in the database, $n$ is the number of variables and $r$ is the number of $W$ variables. This relatively low order of complexity makes LCD2 appropriate for exploring possible causal relationships in databases that contain a very large number of records (on the order of hundreds of thousands) and a moderately large number of measured variables per record (on the order of hundreds).

### Applying the LCD2 Algorithm to the Infant Mortality Database

We implemented LCD2 in the PERL programming language. It takes as input the infant birth and death dataset $D$ and a set of $W$ variables. The $W$ variables we used were *Race of the mother* and *child gen-*

*der.* Race of the mother is determined biologically at the time of conception and none of the variables in $D$ could causally influence it. Child gender is determined randomly at the time of conception and none of the variables in $D$ are known to play a causal role in this. Hence these two variables were assumed not to be caused by any of the other variables in $D$. A default threshold of 0.9 [6] was used for the various dependence and independence tests. It took 60 hours to examine pairwise all the attributes in $D$ and output the nine discovered causes. The program was run on a Gateway computer with a 400 MHz intel processor, 256 megabytes of RAM, running under the Windows NT operating system.

## RESULTS

When applied to the infant birth and death dataset, LCD2 output nine purported causal relationships. Table 2 contains the relationships and Table 3 gives an explanation for the variables in those relationships. Tables 4, 5, and 6 show the probability distributions associated with relationships 1, 3, and 8, respectively, in Table 2.

Table 2: The output of LCD2

| | |
|---|---|
| 1. | MATERNAL_EDUCATION ⇒ DELIVERY_CONDUCTOR |
| 2. | MATERNAL_EDUCATION ⇒ MATERNAL_AGE |
| 3. | MARITAL_STATUS_MOTHER ⇒ DELIVERY_CONDUCTOR |
| 4. | MARITAL_STATUS_MOTHER ⇒ MATERNAL_AGE |
| 5. | PRENATAL_CARE_START ⇒ DELIVERY_FACILITY |
| 6. | PRENATAL_CARE_START ⇒ DELIVERY_CONDUCTOR |
| 7. | PRENATAL_CARE_ADEQUACY ⇒ PRENATAL_CARE_START |
| 8. | BIRTH_WEIGHT ⇒ INFANT_OUTCOME_ONE_YEAR |
| 9. | BIRTH_WEIGHT ⇒ DELIVERY_CONDUCTOR |

Note: The direction of the arrow goes from cause to effect. For all the above causal relationships, the $W$ (instrumental) variable was *maternal race.*

Table 3: Variables and what they signify

| Variable Name | Explanation |
|---|---|
| INFANT_OUTCOME_ONE_YEAR | If child was alive at first birthday |
| MATERNAL_EDUCATION | Years of education of the mother |
| DELIVERY_CONDUCTOR | Care giver conducting the delivery |
| MATERNAL_AGE | Age of mother at delivery |
| MARITAL_STATUS_MOTHER | Marital status of the mother |
| PRENATAL_CARE_START | Trimester prenatal care began |
| DELIVERY_FACILITY | Place or facility of delivery |
| PRENATAL_CARE_ADEQUACY | Adequacy of care recode* |
| BIRTH_WEIGHT | Weight of the infant at birth |
| MATERNAL_RACE | Race of mother |

* This code is based on a modified Kessner criterion. Month prenatal care began, number of prenatal care visits and gestational period are the items used to generate this.

Table 4: Conditional probability table of delivery care giver given maternal education

| Maternal Education | Delivery care giver | | | | |
|---|---|---|---|---|---|
| | MD | DO | CNM | OM | Other |
| 0–8 years | 0.868* | 0.028 | 0.064 | 0.018 | 0.022 |
| 9–11 years | 0.888 | 0.041 | 0.056 | 0.004 | 0.011 |
| 12 years | 0.910 | 0.041 | 0.040 | 0.002 | 0.007 |
| 13–15 years | 0.927 | 0.030 | 0.036 | 0.002 | 0.006 |
| 15+ years | 0.940 | 0.022 | 0.031 | 0.003 | 0.005 |

* The probability that Delivery Conductor is an MD *given* that Maternal education is less than nine years.
MD — Doctor of Medicine; DO — Doctor of Osteopathy; CNM — Certified Nurse Midwife, OM— Other Midwife.

Table 5: Conditional probability table of delivery conductor given marital status of mother

| Marital Status | Delivery Conductor | | | | |
|---|---|---|---|---|---|
| | MD | DO | CNM | OM | Other |
| Married | 0.920* | 0.032 | 0.037 | 0.005 | 0.006 |
| Unmarried | 0.892 | 0.040 | 0.056 | 0.002 | 0.012 |

* The probability that Delivery Conductor is an MD *given* that Mother *is* married.
MD — Doctor of Medicine; DO — Doctor of Osteopathy; CNM — Certified Nurse Midwife, OM — Other Midwife.

## DISCUSSION AND CONCLUSION

In this section we discuss the biological plausibility [11] of the LCD2 output. We realize that additional evaluation is needed, and as stated in the next section, we intend to pursue it.

Out of the nine relationships in Table 2, eight appear plausible. Due to space limitations we will only elaborate a subset of those relationships. Causal relationship #7 linking *the adequacy of prenatal care* to *start month of prenatal care* seems equivocal. The adequacy of prenatal care was derived in the dataset as a function of when prenatal care began, number of prenatal visits and duration of the gestational period; thus, in a sense there is a causal relationship, but it is more definitional than real. Causal relationship #1 postulates that *maternal education* is a cause of *delivery conductor.* Education is an influential component of the socio-economic status of an individual that has a bearing on access to good health care. With in-

Table 6: Conditional probability table of infant outcome given infant birth weight

| Birth Weight | Infant outcome at one year | |
|---|---|---|
| | Survived | Died |
| <1500 gms. | 0.713* | 0.287 |
| 1500–2499 gms. | 0.977 | 0.023 |
| ≥ 2500 gms. | 0.997 | 0.003 |

*The probability that Infant outcome at one year equals Survived *given* that Infant Birth Weight is <1500 grams.

creasing education, the chances of obtaining health insurance and access to better health care may improve. Having the delivery conducted by an MD indicates this enhanced access. Table 4 shows that as the years of maternal education increases, the probability of delivery conductor being an MD increases. Maternal education as a causal factor of Delivery conductor is brought out in this relationship. Causal relationship #3 proposes *mother's marital status* as a causal factor in the choice of delivery conductor. Teenage mothers are likely to be unmarried and have reduced access to good health care. Marriage may improve socio-economic status resulting in a better choice of health care provider. Table 5 shows that for married mothers the probability of delivery conductor being an MD is higher. Marital status as a causal factor of Delivery Conductor is brought out in this relationship. Causal relationship #8 from *Birth weight* to *Infant outcome at one year* turns out to be interesting and well-documented in literature [2], [12]. From Table 6 we see that as the birth weight increases from less than 1500 grams to 1500–2499 grams and then to 2500 or more grams, the probability of survival increases from 0.713 to 0.977 to 0.997.

In summary, the LCD2 algorithm appears to be outputting relationships that on the whole are plausibly causal. None of the relationships found thus far, however, is clinically novel. Nevertheless, the output of LCD2 could be useful to focus additional consideration and study of causal relationships of interest. In the next section, we outline how we plan to apply the algorithm to search further for novel relationships that are clinically useful.

## Future Research

We plan to evaluate the causal relationships output by LCD2 by giving OB/GYN clinicians these relationships and asking them to rate them in terms of causal plausibility. Following an evaluation design in [13], we will intersperse randomly generated relationships among the output of LCD2, so that the clinicians will not be biased by knowledge of the origin of the relationships.

We also plan to re-implement LCD2 in C++ to improve its efficiency. We will apply this faster version of the algorithm to the full infant birth and death database, consisting of approximately four million records.

We used a threshold of 0.9 in the tests of dependence and independence performed by LCD2. We plan to experiment with lowering this threshold. By doing so, we will increase the number of relationships output by LCD2. The false positive rate is likely to increase as well. We plan to study the six tests underlying LCD2 and their particular significant levels in order to gain insight into when and why the algorithm fails and what might be done to improve it.

**Acknowledgements**

This work was supported in part by the National Library of Medicine training grant LM07059 and R01-LM06696.

## References

[1] Wesley C. Salmon. *Causality and Explanation.* Oxford University Press, New York, 1998.

[2] B. Luke, C. Williams, J. Minogue, and L Keith. The changing pattern of infant mortality in the US: The role of prenatal factors and their obstetrical implications. *International Journal of Gynaecology and Obstetrics*, 40(3):199–212, 1993.

[3] Myron E. Wegman. Annual summary of vital statistics—1992. *Pediatrics*, 92(6):743–754, 1993.

[4] National Center for Health Statistics. *1991 Birth Cohort Linked Birth/Infant Dealth Data Set*, May 1996. CD-ROM Series 20—No. 7.

[5] Judea Pearl. *Probabilistic Reasoning in Intelligent Systems.* Morgan Kaufmann, San Francisco, California, 1991.

[6] Gregory F. Cooper. A simple constraint-based algorithm for efficiently mining observational databases for causal relationships. *Data Mining and Knowledge Discovery*, 1:203–224, 1997.

[7] Kenneth A. Bollen. *Structural Equations With Latent Variables.* Wiley, New York, 1989.

[8] J.H. Fetzer, editor. *Probability and Causality.* D. Reidel Publishing Company, Boston, 1989.

[9] G.F. Cooper and E. Herskovits. A Bayesian method for the induction of probabilistic networks from data. *Machine Learning*, 9:309–347, 1992.

[10] David Heckerman, Dan Geiger, and David M. Chickering. Learning Bayesian networks: The combination of knowledge and statistical data. *Machine Learning*, 20(3):197–243, 1995.

[11] Judith S Mausner and Shira Kramer. The concept of causality and steps in the establishment of causal relationships. In *Epidemiology—An Introductory Text*, Philadelphia, PA, 1985. W.B. Saunders.

[12] D.D. McIntire, S.L. Bloom, B.M. Casey, and K.J. Leveno. Birth weight in relation to morbidity and mortality among newborn infants. *NEJM*, 340(16):1234–1238, 1999.

[13] Peter Spirtes and Greg Cooper. An experiment in causal discovery using a pneumonia database. In *Artificial Intelligence and Statistics 99*, pages 162–168, San Francisco, California, 1999. Morgan Kaufmann Publishers.