JMLR: Workshop and Conference Proceedings 3: 1-33 WCCI2008 workshop on causality

Design and Analysis of the Causation and Prediction Challenge

Isabelle Guyon ISABELLE@CLOPINET.COM
Clopinet, California

Constantin Aliferis CONSTANTIN.ALIFERIS@NYUMC.ORG
New York University, New York

Greg Cooper GFC@PITT.EDU
University of Pittsburgh, Pennsylvania

André Elisseeff AEL@ZURICH.IBM.COM
IBM Research, Zürich

Jean-Philippe Pellet JEP@ZURICH.IBM.COM
IBM Research and ETH, Zürich

Peter Spirtes PS7Z@ANDREW.CMU.EDU
Carnegie Mellon University, Pennsylvania

Alexander Statnikov ALEXANDER.STATNIKOV@VANDERBILT.EDU
Vanderbilt University, Tennessee

Editor: Neil Lawrence

Abstract

We organized for WCCI 2008 a challenge to evaluate causal modeling techniques, focusing on predicting the effect of "interventions" performed by an external agent. Examples of that problem are found in the medical domain to predict the effect of a drug prior to administering it, or in econometrics to predict the effect of a new policy prior to issuing it. We concentrate on a given target variable to be predicted (*e.g.,* health status of a patient) from a number of candidate predictive variables or "features" (*e.g.,* risk factors in the medical domain). Under interventions, variable predictive power and causality are tied together. For instance, both smoking and coughing may be predictive of lung cancer (the target) in the absence of external intervention; however, prohibiting smoking (a possible cause) may prevent lung cancer, but administering a cough medicine to stop coughing (a possible consequence) would not. We propose four tasks from various application domains, each dataset including a training set drawn from a "natural" distribution in which no variable are externally manipulated and three test sets: one from the same distribution as the training set and two corresponding to data drawn when an external agent is manipulating certain variables. The goal is to predict a binary target variable, whose values on test data are withheld. The participants were asked to provide predictions of the target variable on test data and the list of variables (features) used to make predictions. The challenge platform remains open for post-challenge submissions and the organization of other events is under way (see http://clopinet.com/causality).

Keywords: challenge, competition, causality, causal discovery, feature selection, intervention, manipulation.

1. Introduction

The problem of attributing causes to effects is pervasive in science, medicine, economics and almost every aspect of our everyday life involving human reasoning and decision making. One important goal of causal modeling is to unravel enough of the data generating process to be able to make predictions under manipulations of the system of interest by an external agent (*e.g.,* experiments). Being able to predict the results of actual or potential experiments (consequences or effects)[1] is very useful because experiments are often costly and sometimes impossible or unethical to perform. For instance, in policy-making, one may want to predict "the effect on a population's health status" of "forbidding individuals to smoke in public places" before passing a law. This example illustrates the case of an experiment which is possible, but expensive. On the other hand, forcing people to smoke would constitute an unethical experiment.

The need for assisting policy making and the availability of massive amounts of "observational" data has prompted the proliferation of proposed causal discovery techniques. These techniques estimate the structure of the data generating process from which the effect of intervention can be estimated. Each scientific discipline has its favorite approach (*e.g.,* Bayesian networks in biology and structural equation modeling in the social sciences), not necessarily reflecting a better match of techniques to domains, but rather the historical tradition. Standard benchmarks are needed to foster scientific progress. In organizing a challenge for WCCI on the theme of causality, our goals included:

- Stimulating the causal discovery community to make progress by exposing it to large datasets, whose size is more typical of data mining and machine learning tasks than causal learning.
- Drawing the attention of the computational intelligence community to the importance of causal modeling and discovery problems and the opportunities to explore machine learning and data mining techniques.
- Pointing out possible limitations of current methods on some particularly difficult problems.

The last item is especially relevant for feature selection algorithms emanating from machine learning as most current machne learning methods do not attempt to uncover cause-effect relationships between features and target. This is justified for a prediction task where training and tests sets are obtained by drawing samples identically and independently from the same "natural" distribution. We call this a purely "observational" setting. In that setting, statistical predictive models do not need to model data generative mechanisms and both causal and consequential features may be predictive of a certain target variable. For instance both smoking and coughing are predictive of respiratory disease; one is a cause and the other a symptom (consequence). In contrast, in this challenge, we investigated a setting in which *the training and test data are not necessarily identically distributed.* Test data may be drawn from a post-manipulation distribution that is distinct from the unmanipulated "natural" distribution from which training data are drawn. This problem is related to the more general problem of "distribution shift" or "covariate shift", which has

1. In this paper, we will use interchangeably "manipulation" or "intervention" and "consequence" or "effect".

recently gained the attention of the machine learning community and was the object of a challenge (Quiñonero Candela et al., 2007). In the particular case we are interested in, the post-manipulation distribution results from **actions** or **interventions** of an external agent who is forcing some variables to assume particular values rather than letting the data generative system produce values according to its own dynamics. Acting on a cause of en event can change the event, but acting on a consequence cannot. For instance, acting on a cause of disease like smoking can change the disease state, but acting on the symptom (coughing) cannot. Thus it is extremely important to distinguish between causes and effects to predict the consequences of actions on a given target variable.

The main objective of the challenge was to predict a binary target variable (classification problem) from a set of candidate predictive variables, which may be binary or continuous. For each task of the challenge (*e.g.,* REGED, SIDO, etc.), we have a single training set, but several test sets (associated with the dataset name, *e.g.,* REGED0, REGED1, and REGED2). The training data come from a so-called "natural distribution", and the test data in version zero of the task (*e.g.,* REGED0) are also drawn from the same distribution. We call this test set a "natural" or "unmanipulated" test set. The test data from the two other versions of the task (*e.g.,* REGED1 and REGED2) are "manipulated" test sets resulting from **interventions** of an external agent, which has "manipulated" some or all the variables in some way (excluding the "target" or "response variable"). The effect of such manipulations is to *disconnect the manipulated variables from their natural causes.* This may affect the predictive power of a number of variables in the system, including the manipulated variables. Hence, to obtain optimum predictions of the target variable, feature selection strategies should take into account such manipulations.

In this challenge, we are focusing on causal relationships between random variables, as opposed to causal relationships between events or objects. We consider only stationary systems in equilibrium, hence eliminating the need for an explicit reference to time in our samples. This setup is typical of so-called "cross-sectional" studies in medicine (as opposed to "longitudinal" studies). In practice, this means that the samples for each version of the test set, *e.g.,* REGED0, REGED1, and REGED2, are drawn independently, according to a given distribution, which changes only between test set version. Having no explicit reference to time may be surprising to researchers new to causal modeling, since causes must always precede their effects. Causal models in this context enforce an order of evaluation of the variables, without reference to an exact timing.[2]

The type of causal relationships under consideration have often been modeled as Bayesian causal networks or structural equation models (SEM) (Pearl, 2000; Spirtes et al., 2000; Neapolitan, 2003). In the graphical representation of such models, an arrow between two variables $A \rightarrow B$ indicates the direction of a causal relationship: A causes B. A node in of the graph, labeled with a particular variable X, represents a mechanism to evaluate the value of X given the parent node variable values. For Bayesian networks, such evaluation is carried out by a conditional probability distribution $P(X|Parents(X))$ while for structural equation models it is carried out by a function of the parent variables, plus some noise. Learning a causal graph can be thought of as a model selection problem: Alternative graph

2. When manipulations are performed, we must specify whether we sample from the distribution before or after the effects of the manipulation have propagated. Here we assume that we sample after the effects have propagated.

architectures are considered and a selection is performed, either by ranking the architectures with a global score (*e.g.,* a marginal likelihood, or a penalty-based cost function), or by retaining only graphs that fulfill a number of constraints such as dependencies or independencies between subsets of variables.

Bayesian networks and SEMs provide a convenient language to talk about the type of problem we are interested in, but our setting does not preclude of any particular model. Some of the data used in the challenge were generated by real unknown processes, which probably violate some commonly made causal modeling assumptions, such as "causal sufficiency"[3], linearity, Gaussian noise, absence of cycles, etc. By adopting a predictive modeling perspective, we purposely took some distance with the interpretation of causal models as data generative models. The goal of the challenge was not to reverse engineer the data generative process, it is to make accurate predictions of a target variable. To sharpen this distinction, we made available only a limited amount of training data, such that the learner may not necessarily be able to reliably determine all conditional dependencies and independencies. Hence, modeling strategies making radical simplifying assumptions might do better than strategies trying to be faithful to the data generative process, because of the well-known fit *vs.* robustness (or bias *vs.* variance) tradeoff.

2. General setting

We created a web site from which data and instructions on how to participate were outlined: http://clopinet.com/causality. This first causality challenge is part of a larger program, which we initiated, called the "causality workbench"; the web site hosts repositories of code, data, models, publications and other events, including challenges and teleconference seminars. Our first challenge started on December 15, 2007 and ended on April 30, 2008. Four datasets were proposed and progressively introduced (the last one being released 2 months prior the end of the challenge). More details on the datasets are found in Section 3.

Our challenge is formatted in a similar way to most machine learning problems: pairs of training examples $\{\mathbf{x}, y\}$ are provided. The goal is to predict the target variable y for new test instances of $\mathbf{x}$. The elements of vector $\mathbf{x}$ are interchangeably called "variables" or "features" in this paper. Unlike most machine learning problems, the training and test sets are not always distributed similarly. We provide large test sets to obtain statistically significant results. Both the training and the unlabeled test sets were provided from the beginning of the competition. We required that the participants would not use the unlabeled test data to train their models, and this rule was enforced by verifying the code of the best ranking entrants after the end of the challenge (see Appendix B). This rule was motivated by several considerations: (1) We are investigating problems in which only "observational" training data are available for model building. Test data are not supposed to be available at model building time; we use them only to test the ability of our model to make predictions about the effect of hypothetical actions performed on the system in the future. (2) In a challenge, we need very large test sets to obtain small error bars on the participant performances, otherwise most differences between algorithms would not be statistically

3. "Causal sufficiency" roughly means that there are no unobserved common causes of the observed variables.

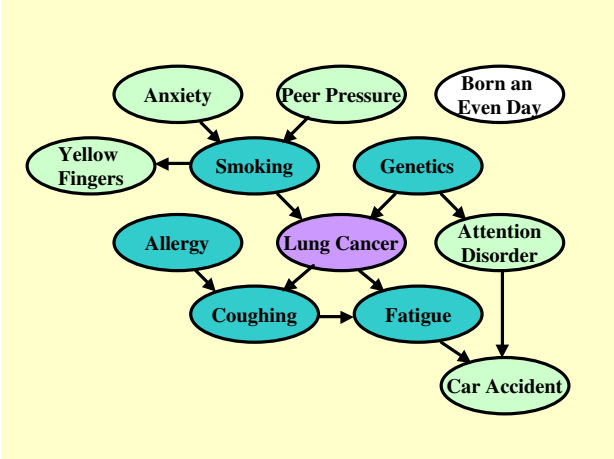
significant. However, such large amount of "manipulated" test data would not be available all at once in many real world situations.

Prediction results and features sets could be submitted on-line to get immediate feed-back, as a means of stimulating participation. To limit the amount of knowledge that could be gained from viewing test set results, the participants were only informed about the quartile of their method's performances. In previous challenges we organized (Guyon et al., 2006a,b, 2008b), we provided feed-back on a small validation set, whose target values were released shortly before the end of the challenge, and we used a separate larger test to perform the final evaluation. In this challenge, we developed this new way of providing feed-back (using performance quartiles) because information about the post-manipulation distribution (distinct from the training data "natural" distribution) could be induced from a more detailed form of performance feed-back on a validation set. The quartile method achieves essentially the same goal of stimulating the participants while simplifying the challenge protocol.

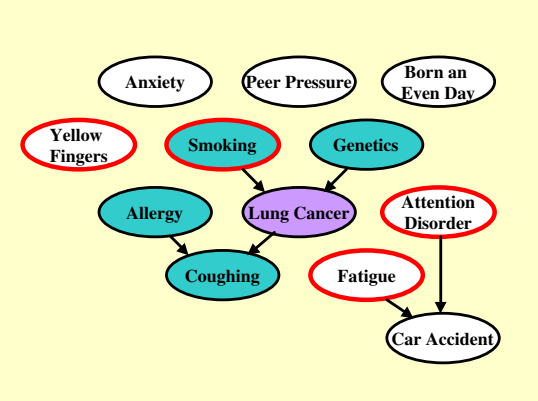
Another difference compared to our previous challenges is that we did not request that the participants return results on all tasks of the challenge. For each task, they were only required to return predictions on all three versions of any given test set (manipulated or not). In this way, we intended to lower the level of effort of participation because we knew many algorithms lend themselves only to certain kinds of data. To encourage participants to submit results on more than one task, we set up an exponential reward system: a prize of $100 was promised for winning on any of the 4 tasks, but the progression of the rewards for winning on 2, 3, or 4 datasets was $400, $900, and $1600. This successfully encouraged entrants to submit on all datasets. Another final difference from previous challenges is that we authorized only one final entry (as opposed to 5 in previous challenges) to compensate for the fact that participants had 4 chances of winning (one for each dataset). In this way, we limited the statistical risk that the winning entry be better only "by chance". However, we did allow submissions of multiple prediction results for *nested subsets of variables*, with the purpose of obtaining performance curves as a function of number of features. In Section 5, our initial analysis is based on the best result in the performance curve for each participant. We complemented it by an analysis making pairwise comparisons of entries at the same number of features, to account for a possible bias detrimental to the participants who provided single predictions.

To introduce the participants to the problem of making predictions under interventions, we provided a tutorial (Guyon et al., 2007), and we created a toy example, which was not part of the challenge, but which was interfaced to the challenge platform in the same way as the other datasets. The participants could use it for practice purposes, and we provided guidance on how to solve the problem on the web site. We briefly describe this example, illustrated in Figure 1, to clarify the challenge. More details are found on the website of the challenge.

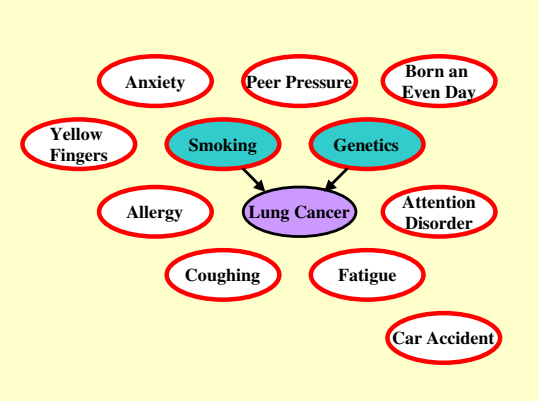
LUCAS0: The toy example of Figure 1-a models the problem of predicting lung cancer as a causal network. Each node represents a variable/feature and the arcs represent causal relationships, *i.e.,* $A \rightarrow B$ represents that A is a cause of B. The target variable is "Lung Cancer". Each node in the graph is associated with a table of conditional probabilities $P(X = x|Parent_1(X) = p_1, Parent_2(X) = p_2, ...)$ defining the "natural" distribution. The generative model is a Markov process (a so-called "Bayesian network"), so the state of



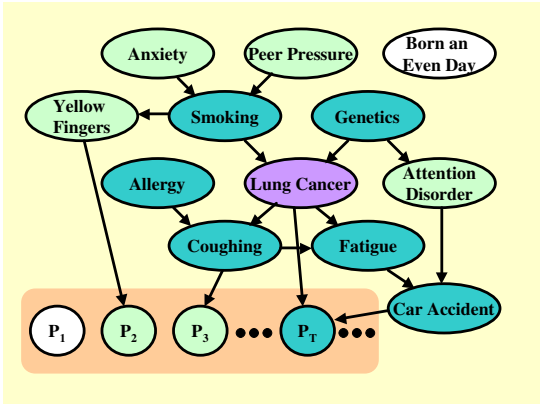
(a) LUCAS0



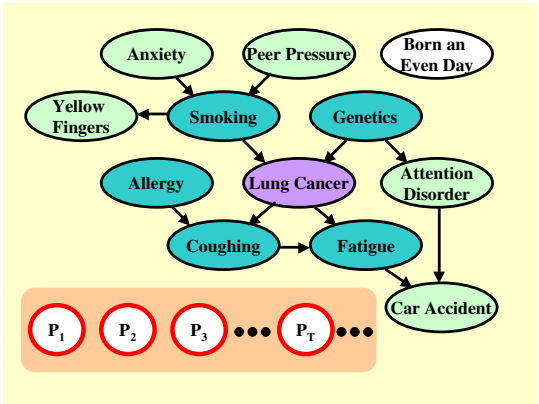
(b) LUCAS1



(c) LUCAS2



(d) LUCAP0



(e) LUCAP1

Figure 1: **Lung cancer toy example.** The dark green nodes represents the minimal Markov blanket or "Markov boundary" (MB) of the target variable "Lung Cancer". The white nodes are independent of the target. Given the MB, both white and light green nodes are (conditionally) independent of the target. The manipulated nodes are emphasized in red. As a result of being manipulated, they are disconnected from their original causes and the MB is restricted to the remaining dark green nodes. See text.

the children is stochastically determined by the states of the parents. The values must be drawn in a certain order, so that the children are evaluated after their parents. Both the training and test sets of LUCAS0 are drawn according the natural distribution. In the figure, we outline in dark green the Markov blanket of the target, which includes all targets's parents (node immediate antecedents), children (node immediate descendants), and spouses (immediate antecedents of an immediate descendant). The Markov blanket (MB) is the set of variables such that the target is independent of all other variables given MB.[4] It is widely believed that, if the MB were perfectly known, adding more variables to the feature set would be unnecessary to make optimal predictions of the target variable. However, this statement depends on the criterion of optimality and is true only in the sample limit and if the predictor is asymptotically unbiased (Tsamardinos and Aliferis, 2003). For example, a linear classifier may benefit from the inclusion of non-MB features, even in the sample limit and with perfect knowledge of the MB, if the functional relation of the target and the MB is non-linear. In this challenge, the goal is not to discover the MB, it is to make best predictions of the target variable on test data.

LUCAS1: In the example of Figure 1-b, the training data are the same as in LUCAS0. We model a scenario in which an external agent manipulates some of the variables of the system, circled in red in the figure (Yellow Fingers, Smoking, Fatigue, and Attention Disorder). The intention of such manipulations may include disease prevention or cure. The external agent sets the manipulated variables to desired values, hence "disconnecting" those variables from their parents. The other variables are obtained by letting the system evolve according to its own dynamics. As a result of manipulations, many variables may become disconnected from the target and the Markov blanket (MB) may change. If the identity of the manipulated variables is revealed (as in the case of REGED1 and MARTI1), one can deduce from the graph of the natural distribution inferred from training data which variables to exclude from the set of predictive variables. In particular, the MB of the post-manipulation distribution is a restriction of the MB of the natural distribution resulting from the removal of manipulated children and spouses whose children are all manipulated (unless it is also a parent of the target).

LUCAS2: In Figure 1-c we manipulated all the variables except the target. As a result, only the direct causes of the target are predictive, and they coincide with the Markov blanket (MB) of the post-manipulation distribution.

LUCAP0: In Figure 1-d, we are modeling the following situation: Imagine that we have REAL data generated from some UNKNOWN process (we do not know the causal relationships among variables). Further, for various reasons, which may include practical reasons, ethical reasons, or cost, we are unable to carry out any kind of manipulation on the real variables, so we must resort to performing causal discovery and evaluating the effectiveness of our causal discovery using *unmanipulated* data (data drawn from the natural distribution). To that end, we add a large number of artificial variables called "probes", which are generated from some functions (plus some noise) of subsets of the real variables. We shuffle the order of all the variables and probes not to make it too easy to

4. Other definitions of the Markov blanket are possible. Our definition coincides with what other authors call Markov boundary or "minimal" Markov blanket. Although we refer to "the" Markov blanket, for some distributions it is not unique and it does not always coincide with the sets of parents, children and spouses. But we limit ourselves to this case in the example, for simplicity.

identify the probes. For the probes we (the organizers) have perfect knowledge of the causal relationships. For the other variables, we only know that some of them (light green nodes) might be predictive while not belonging to the MB, and some of them (dark green nodes) might belong to the MB. The members of the MB include some real variables and some probes. To assess feature selection methods, we use the probes by computing statistics such as the fraction of non-MB probes in the feature subset selected.

LUCAP1 and LUCAP2: While we cannot manipulate the real variables in our model setup, we can manipulate the probes. The probe method allows us to conservatively evaluate causal feature selection algorithms, because we know that the output of an algorithm should not include any probe for a distribution where all probes are manipulated. The test sets of LUCAP1 and LUCAP2 (Figure 1-e) are obtained by manipulating all probes (in every sample) in two different ways. The training data are the same as in LUCAP0. Knowing that we manipulated all probes, and that probes can only be non-causes of the target, a possible strategy is to select only features that are causes of the target.[5] If this strategy is followed, the fraction of probes in the feature set selected allows us to compute an estimate the fraction of non-causes wrongly selected.[6]

3. Description of the datasets

We use two types of data:

- **Re-simulated data:** We train a "causal" model (a causal Bayesian network or a structural equation model) with real data. The model is then used to generate artificial training and test data for the challenge. Truth values of causal relationships are known for the data generating model and used for scoring causal discovery results. REGED is an example of re-simulated dataset.
- **Real data with probe variables:** We use a dataset of real samples. Some of the variables may be causally related to the target and some may be predictive but non-causal. The nature of the causal relationships of the variables to the target is unknown (although domain knowledge may allow us to validate the discoveries to some extent). We have added to the set of real variables a number of distractor variables called "probes", which are generated by an artificial stochastic process, including explicit functions of some of the real variables, other artificial variables, and/or the target. All probes are non-causes of the target, some are completely unrelated to the target. The identity of the probes in concealed. The fact that truth values of causal relationships are known only for the probes affects the evaluation of causal discovery, which is less reliable than for artificial data.

The **training data** and test sets labeled 0 are generated from a so-called "natural" pre-manipulation distribution. The variable values are sampled from the system when it is allowed to evolve according to its own dynamics, after it has settled in a steady state. For the probe method, the system includes the artificial probe generating mechanism. **Test**

5. Note however that some of the real variables that are non-causes may be predictive, so eliminating all non-causes of the target is a sure way to eliminate all probes but not necessarily an optimum strategy.
6. The validity of the estimation depends on many factors, including the number of probes and the distributional assumptions of non-causes made in the probe data generative process.

Table 1: **Datasets.** All target variables are binary. Each dataset has three test sets of the same size numbered 0, 1, and 2.

Dataset	Domain	Type	Features	Feat. #	Train #	Test #
REGED	Genomics	Re-simulated	Numeric	999	500	20000
SIDO	Pharmacology	Real + probes	Binary	4932	12678	10000
CINA	Econometrics	Real + probes	Mixed	132	16033	10000
MARTI	Genomics	Re-simulated	Numeric	999	500	20000

sets labeled 1 and 2 are generated from a so-called **post-manipulation distribution**. An external agent performs an "intervention" on the system. Depending on the problem at hand, interventions can be of several kinds, *e.g.,* clamping one or several variables to given values or drawing them from an alternative distribution, then sampling the other variables according to the original conditional probabilities. In our design, **the target variable is never manipulated**. For the probe method, since we do not have the possibility of manipulating the real variables, we only manipulate the probes. The effect of manipulations is to disconnect the variables from their natural causes. Manipulations allow us to eventually influence the target, if we manipulate causes of the target. Manipulating non-causes should have no effect on the target. Hence, without inferring causal relationships, it should be more difficult to make predictions for post-manipulation distributions.

We proposed four tasks (Table 1):

REGED (REsimulated Gene Expression Dataset): Find genes which could be responsible for lung cancer. The data are "re-simulated", *i.e.,* generated by a model derived from real human lung-cancer microarray gene expression data. From the causal discovery point of view, it is important to separate genes whose activity causes lung cancer from those whose activity is a consequence of the disease. All three datasets (REGED0, REGED1, and REGED2) include 999 features (no hidden variables or missing data), the same 500 training examples, and different test sets of 20000 examples. The target variable is binary; it separates malignant samples (adenocarcinoma) from control samples (squamous cells). The three test sets differ in their distribution. REGED0: No manipulation (distribution identical to the training data). REGED1: Variables in a given set are manipulated and their identity is disclosed. REGED2: Many variables are manipulated, including all the consequences of the target, but the identity of the manipulated variables was not disclosed. When variables are manipulated, the model is allowed to evolve according to its own mechanism until the effect of the manipulations propagate.

SIDO (SImple Drug Operation mechanisms) contains descriptors of molecules which have been tested against the AIDS HIV virus. The target values indicate the molecular activity (+1 active, −1 inactive). The causal discovery task is to uncover causes of molecular activity among the molecule descriptors. This would help chemists in the design of new compounds, retaining activity, but having perhaps other desirable properties (less toxic, easier to administer). The molecular descriptors were generated programmatically from the three dimensional description of the molecule, with several programs used by pharmaceutical companies for QSAR studies (Quantitative Structure-Activity Relationship). For example, a descriptor may be the number of carbon molecules, the presence of an aliphatic cycle, the length of the longest saturated chain, etc. The dataset includes 4932 variables (other

than the target), which are either molecular descriptors (all potential causes of the target) or "probes" (artificially generated variables that are not causes of the target). The training set and the unmanipulated test set SIDO0 are similarly distributed. They are constructed such that some of the "probes" are effects (consequences) of the target and/or of other real variables, and some are unrelated to the target or other real variables. Hence, both in the training set and the unmanipulated test set, all the probes are non-causes of the target, yet some of them may be "observationally" predictive of the target. In the manipulated test sets SIDO1 and SIDO2, all the "probes" are manipulated in every sample by an external agent (*i.e.,* set to given values, not affected by the dynamics of the system) and can therefore not be relied upon to predict the target. The identity of the probes is concealed. They are used to assess the effectiveness of the algorithms to dismiss non-causes of the target for making predictions in manipulated test data. In SIDO1, the manipulation consists in a simple randomization of the variable values, whereas in SIDO2 the values are chosen to bias prediction results unfavorably, if the manipulated variables are chosen as predictors (adversarial design).

CINA (Census Is Not Adult) is derived from census data (the UCI machine-learning repository Adult database). The data consists of census records for a number of individuals. The causal discovery task is to uncover the socio-economic factors affecting higher income (the target value indicates whether the income exceeds 50K). The 14 original attributes (features) including age, workclass, education, marital status, occupation, native country, etc. are continuous, binary, or categorical. Categorical variables were converted to multiple binary variables (as we shall see, this preprocessing, which facilitates the tasks of some classifiers, complicates causal discovery). Distracter features or "probes" (artificially generated variables, which are not causes of the target) were added. In training data, some of the probes are effects (consequences) of the target and/or of other real variables. Some are unrelated to the target or other real variables. Hence, some of the probes may be correlated to the target in training data, although they do not cause it. The unmanipulated test data in CINA0 are distributed like the training data. Hence, both causes and consequences of the target might be predictive in the unmanipulated test data. In contrast, in the manipulated test data of CINA1 and CINA2, all the probes are manipulated by an external agent (*i.e.,* set to given values, not affected by the dynamics of the system) and therefore they cannot be relied upon to predict the target. In a similar way to SIDO, the difference between versions 1 and 2 is that in version 1 the probe values are simply randomized whereas in version 2 they are chosen in an adversarial way.

MARTI (Measurement ARTIfact) is obtained from the same data generative process as REGED, a source of simulated genomic data. Similarly to REGED the data do not have hidden variables or missing data, but a noise model was added to simulate the imperfections of the measurement device. The goal is still to find genes, which could be responsible of lung cancer. The target variable is binary; it indicates malignant samples *vs.* control samples. The feature values representing measurements of gene expression levels are assumed to have been recorded from a two-dimensional microarray 32x32. The training set was perturbed by a zero-mean correlated noise model. The test sets have no added noise. This situation simulates a case where we would be using different instruments at "training time" and "test time", *e.g.,* we would use DNA microarrays to collect training data and PCR for testing. We avoided adding noise to the test set because it would be too difficult to filter it without

visualizing the test data or computing statistics on the test data, which we forbid. So the scenario is that the second instrument (used at test time) is more accurate. In practice, the measurements would also probably be more expensive, so part of the goals of training would be to reduce the size of the feature set (we did not make this a focus in this first challenge).

The problems proposed are challenging in several respects:

- Several assumptions commonly made in causal discovery are violated, including "causal sufficiency"[7], "faithfulness"[8], "linearity", and "Gaussianity".
- Relatively small training sets are provided, making it difficult to infer conditional independencies and learning distributions.
- Large numbers of variables are provided, a particular hurdle for some causal discovery algorithms that do not scale up.

More details on the datasets, including the origin of the raw data, their preparation, past usage, and baseline results can be found in a Technical Report (Guyon et al., 2008).

4. Evaluation

The participants were asked to return *prediction scores* or **discriminant values** v for the target variable on test examples, and a **list of features** used for computing the prediction scores, sorted in order of decreasing predictive power, or unsorted. The classification decision is made by setting a threshold θ on the discriminant value v: predict the positive class if $v > \theta$ and the negative class otherwise. The participants could optionally provide results for nested subsets of features, varying the subset size by powers of 2 (1, 2, 4, 8, etc.).

Tscore: The participants were ranked according to the area under the ROC curve (AUC) computed for test examples (referred to as Tscore), that is the area under the curve plotting sensitivity *vs.* (1− specificity) when the threshold θ is varied (or equivalently the area under the curve plotting sensitivity *vs.* specificity). We call "sensitivity" the error rate of the positive class and "specificity" the error rate of the negative class. The AUC is a standard metric in classification. If results were provided for nested subsets of features, the best Tscore was retained. There are several ways of estimating error bars for the AUC. We use a simple heuristic, which gives us approximate error bars, and is fast and easy to implement: we find on the AUC curve the point corresponding to the largest balanced accuracy BAC = 0.5 (sensitivity + specificity). We then estimate the standard deviation of the BAC as:

$$\sigma = \frac{1}{2}\sqrt{\frac{p_+(1-p_+)}{m_+} + \frac{p_-(1-p_-)}{m_-}}\ , \tag{1}$$

where m_+ is the number of examples of the positive class, m_- is the number of examples of the negative class, and p_+ and p_- are the probabilities of error on examples of the positive and negative class, approximated by their empirical estimates, the sensitivity and the specificity (Guyon et al., 2006b).

7. "Causal sufficiency" roughly means that there are no unobserved common causes of the observed variables.
8. "Faithfulness" roughly means that every conditional independence relation that holds in the population is entailed to hold for all values of the free parameters.

Table 2: **Best scores of ranked entrants.** The table shows the results of the best entries of the ranked entrants and their corresponding scores: Top Tscore = area under the ROC curve on test data for the top ranked entries; Top Fscore = a measure of "causal relevance" of the features used during the challenge (see text). For comparison, we also include the largest reachable score, which was obtained by including reference entries made by the organizers using knowledge about the true causal relationships (Max Ts and Max Fs).

Dataset	Top Tscore		Max Ts	Top Fscore		Max Fs
REGED0	Yin-Wen Chang	1.000±0.001	1.000	Gavin Cawley	0.941±0.036	1.000
REGED1	Marius Popescu	0.989±0.003	0.998	Yin-Wen Chang	0.857±0.062	1.000
REGED2	Yin-Wen Chang	0.839±0.005	0.953	CaMML Team	1.000±0.153	1.000
SIDO0	J. Yin & Z. Geng Gr.	0.944±0.008	0.947	H. Jair Escalante	0.844±0.007	1.000
SIDO1	Gavin Cawley	0.753±0.014	0.789	Mehreen Saeed	0.724±0.007	1.000
SIDO2	Gavin Cawley	0.668±0.013	0.767	Mehreen Saeed	0.724±0.007	1.000
CINA0	Vladimir Nikulin	0.976±0.003	0.979	H. Jair Escalante	0.955±0.032	1.000
CINA1	Gavin Cawley	0.869±0.005	0.898	Mehreen Saeed	0.786±0.039	1.000
CINA2	Yin-Wen Chang	0.816±0.005	0.891	Mehreen Saeed	0.786±0.039	1.000
MARTI0	Gavin Cawley	1.000±0.001	1.000	Gavin Cawley	0.870±0.048	1.000
MARTI1	Gavin Cawley	0.947±0.004	0.954	Gavin Cawley	0.806±0.063	1.000
MARTI2	Gavin Cawley	0.798±0.006	0.827	Gavin Cawley	0.996±0.153	1.000

Fscore: We also computed other statistics, which were not used to rank participants, but used in the analysis of the results. Those included the number of features used by the participants called "Fnum", and a statistic assessing the quality of causal discovery in the feature set selected called "Fscore". As with the Tscore, we provided quartile feed-back on Fnum and Fscore during the competition. For the Fscore, we used the AUC for the problem of separating features belonging to the Markov blanket of the test set distribution *vs.* other features. Details are provided on the web site of the challenge. As it turns out, for reasons explained in Section 5, this statistic correlates poorly with the Tscore and, after experimenting with various scores, we found better alternatives.

5. Result Analysis

5.1 Best challenge results

We declared three winners of the challenge:

- **Gavin Cawley** (University of East Anglia, UK): Best prediction accuracy on SIDO and MARTI, using Causal explorer and linear ridge regression ensembles. Prize: $400.
- **Yin Wen Chang** (National Taiwan University): Best prediction accuracy on REGED and CINA, using SVM. Prize: $400.
- **Jianxin Yin and Zhi Geng's group** (Peking University, Beijing, China): Best overall contribution, using Partial Orientation and Local Structural Learning (new original causal discovery algorithm and best on Pareto front causation/prediction, *i.e.,* with smallest Euclidian distance to the extreme point with zero error and zero features). Prize: free WCCI 2008 registration.

The top-ranking results are summarized in Table 2. These results are taken from the last entries of the ranked entrants.[9]

Following the rules of the challenge, the participants were allowed to turn in *multiple prediction results* corresponding to *nested subsets of features*. The best Tscore over all feature set sizes was then retained and the performances were averaged over all three test sets for each task REGED, SIDO, CINA, and MARTI. In this way, we encouraged the participants to rank features rather than select a single feature subset, since feature ranking is of interest for visualization, data understanding, monitoring the tradefoff "number of features"/"prediction performance", and prioritizing potential targets of action. The entries of Gavin Cawley (Cawley, 2008) and Yin Wen Chang and Chih-Jen Lin (Chang and Lin, 2008) made use of this possibility of turning in multiple results. They each won on two datasets and ranked second and third on the two others. Their average Tscore over all tasks is almost identical and better than that of other entrants (see Figure 2-b).

The participants who used nested subsets had an advantage over other participants, not only because they could make multiple submissions and be scored on the basis of the best results, but also because the selection of the best point was made with test data drawn from the post-manipulation distribution, therefore implicitly giving access to information on the post-manipulation distribution. By examining the results and the "Fact Sheets", we noticed that most participants having performed causal discovery opted to return a *single feature subset* while those using non-causal feature selection performed feature ranking and opted to return multiple predictions for *nested subsets of features*, therefore introducing a bias in the results. To compensate for that bias, we made pairwise comparisons between classifiers, at equal number of features (see details in Section 5.2). According to this new method of comparison, Jianxin Yin and Zhi Geng's group obtain the best position of the Pareto front of the Fscore *vs.* Tscore graph (see Figure 2-b). Because of this achievement and the originality of the method that they developed (Yin et al., 2008), they were awarded a prize for "best overall contribution".[10] Also noteworthy was the performances of Vladimir Nikulin (Suncorp, Australia), who ranked second on CINA and fourth on REGED and MARTI in average Tscore, based on predictions made with a single feature subset obtained with the "random subset method" (Nikulin, 2008). His average performances were as good as Jianxin Yin and Zhi Geng's group on average in the pairwise comparison of classifiers (Figure 2-b), even though his average Fscore is significantly lower.

Also worthy of attention are the entries of Marc Boullé and Laura E. Brown & Ioannis Tsamardinos, who did not compete towards the prizes (and therefore were not ranked), identified as M.B. and L.E.B. & Y.T. in the figures.[11] Marc Boullé reached Tscore=0.998 for REGED1 and Laura E. Brown & Ioannis Tsamardinos reached Tscore=0.86 on REGED2. Marc Boullé also reached Tscore=0.979 on CINA0 and Tscore=0.898 on CINA1. The entries of Marc Boullé, using a univariate feature selection method and a naïve Bayes classifier (Boullé, 2007a,b) were best on REGED0 and REGED1 and on CINA0 and CINA1. The

9. This table reports the results published on the web site of the challenge, using the original definition of the Fscore, whose effectiveness to assess causal discovery is questioned in Section 5.2.
10. As explained in Section 5.2, the original Fscore had some limitations. In Figure 2 we plot the new Fscore described in that section.
11. As per his own request, the entries of Marc Boullé (M.B.) were marked as "Reference" entries like those of the organizers and did not count towards wining the prizes; Laura E. Brown and Ioannis Tsamardinos (L.E.B. & Y.T.) could not compete because they are close collaborators of some of the organizers.

entry of Laura E. Brown & Ioannis Tsamardinos on REGED2 is significantly better than anyone else's and they are best on average on REGED. They use a novel structure-based causal discovery method (Brown and Tsamardinos, 2008). Finally, Mehreen Saeed ranked fourth and sixth on SIDO and CINA, using a novel fast method for computing Markov blankets (Saeed, 2008). She achieved the best Fscores on SIDO1&2 and CINA1&2.

All the top-ranking entrants we just mentioned supplied their code to the organizers, who could verify that they complied with all the rules of the challenge and that their results are reproducible (see Appendix B).

5.2 Causation and Prediction

One of the goals of the challenge was to test the efficacy of using causal models to make good predictions under manipulations. In an attempt to quantify the validity of causal models, we defined an Fscore (see Section 2). Our first analysis of the challenge results revealed that this score correlates poorly with the Tscore, measuring prediction accuracy. In particular, many entrants obtained a high Fscore on REGED2 and yet a poor Tscore. In retrospect, this is easily understood. We provide a simple explanation for the case of unsorted feature sets in which, for REGED, the Fscore is $0.5(tp/(tp+fn)+tn/(tn+fp))$, where tp is the number of true positive (correctly selected features), fn false negative, tn true negative, and fp false positive. REGED2 has only 2 causally relevant feature (direct causes) in the *manipulated Markov blanket*; *i.e.,* the Markov blanket of the test set distribution, which is manipulated. Most people included these two features in their feature set and obtained $tp/(tp+fn)=1$. Since the number of irrelevant features is by comparison very large (of the order of 1000), even if the number of wrongly selected features fp is of the order of 10, $tn/(tn+fp)$ is still of the order of 1. The resulting Fscore is therefore close to 1. However, from the point of view of the predictive power of the feature set, including 10 false positive rather than 2 makes a lot of difference. We clearly see that the first Fscore we selected was a bad choice.

Definition of a new Fscore. We ended up using as the new Fscore the **Fmeasure** for REGED and MARTI and the **precision** for SIDO and CINA, after experimenting with various alternative measures inspired by information retrieval, see our justification below. We use the following definitions: precision = $tp/(tp+fp)$, recall = $tp/(tp+fn)$ (also called sensitivity), and Fmeasure = 2 precision recall / (precision + recall). Our explorations indicate that precision, recall, and Fmeasure correlate well with Tscore for artificially generated datasets (REGED and MARTI). The *Fmeasure*, which captures the tradeoff between precision and recall, is a good measure of feature set quality for these datasets. However, recall correlates poorly with Tscore for SIDO and CINA, which are datasets of *real variables* with added *artificial probe variables*. This is because, in such cases, we must resort in approximating the recall by the fraction of *real variables* present in the selected feature set, which can be very different from the true recall (the fraction of truly relevant variables). Hence, if many real variables are irrelevant, a good causal discovery algorithm that eliminates them would get a poor estimated recall. Hence, we can only use *precision* as of feature set quality for those datasets. A plot of the new Fscore

vs. Tscore (Figure 2-a) reveals that a significant correlation of 0.84 is attained (pvalue 2.10^{-19}),[12] when the scores are averaged over all datasets and test set versions.

To factor out the variability due to the choice of the classifier, we asked several participants to train their learning machine on all the feature sets submitted by the participants and we redrew the same graphs. The performances improved or degraded for some participants and some datasets, but on average, the correlation between Tscore and the new Fscore did not change significantly.[13] See the on-line results for details.

Pairwise comparisons. In an effort to remove the bias introduced by selecting the best Tscore for participants who returned multiple prediction results for nested subsets of features, we made pairwise comparisons of entries, using the same number of features. Specifically, if one entry used a single feature subset of size n and the other provided results for nested subsets, we selected for the second entry the Tscore corresponding to n by interpolating between Tscore values for the nested subsets. If both entries used nested feature subsets, we compared them at the median feature subset size used by other entrants. If both entries used a single subset of features, we directly compared their Tscores. For each participant, we counted the fraction of times his Tscore was larger than that of others. We proceeded similarly with the Fscore. Figure 2-b shows the resulting plot. One notices that the performances of the winners by Tscore, Gavin Cawley and Yin-Wen Chang, regress in the pairwise comparison and that Jianxin Yin and Zhi Geng's group, Vladimir Nikulin, and Marc Boullé (M.B.), appear now to have better predictive accuracy. Jianxin Yin and Zhi Geng's group stand out on the Pareto front by achieving also best Fscore.

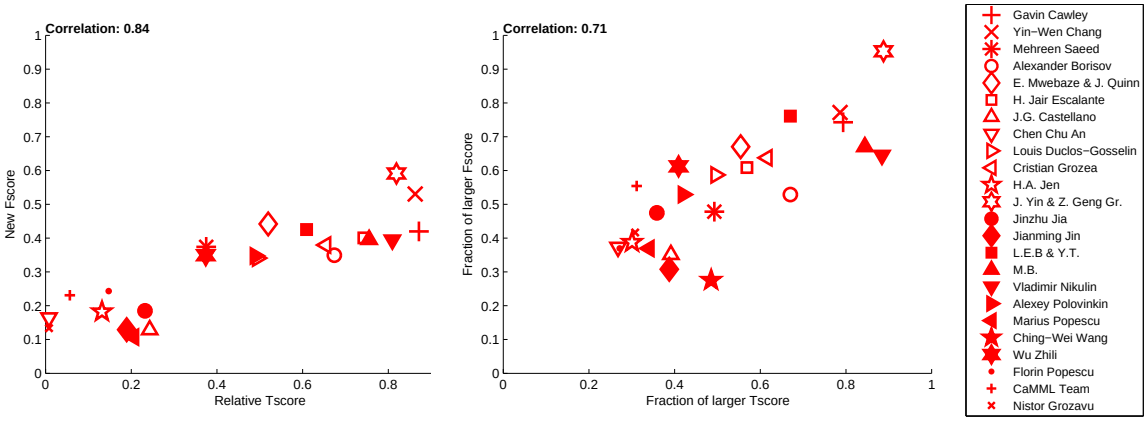
5.3 Methods employed

The methods employed by the top-ranking entrants can be categorized in three families:

- **Causal:** Methods employing causal discovery techniques to unravel cause-effect relationships in the neighborhood of the target.
- **Markov blanket:** Methods for extracting the Markov blanket, without attempting to unravel cause-effect relationships.
- **Feature selection:** Methods for selecting predictive features making no explicit attempt to uncover the Markov blanket or perform causal discovery.

12. This is the pvalue for the hypothesis of no correlation. We use confidence bounds that are based on an asymptotic normal distribution of $0.5 \cdot \log((1+R)/(1-R))$, where R is the Pearson correlation coefficient, as provided by the Matlab statistics toolbox.

13. Computing these scores requires defining truth values for the set of "relevant features" and "irrelevant features". In our original Fscore, we used the Markov blanket of the test set distribution as set of "relevant features". For SIDO and CINA, there is only partial knowledge of the causal graph. The set of "relevant variables" is approximated by all true variables and the probes belonging to the Markov blanket (of the test set distribution). As an additional refinement, we experimented with three possible definitions of "relevant features": (1) the Markov blanket (MB), (2) MB + all causes and effects, and (3) all variables connected to the target through any directed or undirected path. If the test data are manipulated, those sets of variables are restricted to the variables not disconnected from the target as a result of manipulations. We ended up computing the new Fscore for each definition of "relevant features" and performing a weighed average with weights 3, 2, 1. We did not experiment with these weights, but the resulting score correlates better with Tscore than when the Markov blanket of the test distribution alone is used as reference "relevant" feature set. This is an indication that features, which are outside of the Markov blanket may be useful to make predictions (see Section 6 for a discussion).



(a) Regular ranking (b) Pairwise comparisons

Figure 2: **Correlation between feature selection score and prediction.** The new feature selection score Fscore (see text) evaluating the accuracy of causal discovery is plot as a function of the prediction accuracy score Tscore (the area under the ROC curve for test examples). The relative Tscore is defined as (Tscore − 0.5)/(Max Tscore − 0.5). Both Tscore and Fscore are averaged over all tasks and test set versions. (a) Ranking according to the rules of the challenge, selecting the best Tscore for nested feature subset results. (b) Ranking obtained with pairwise comparisons of classifiers, using the same number of features in each comparison.

In this section, we briefly describe prototypical examples of such methods taken among those employed by top-ranking participants.

Causal discovery: The top-ranking entrants who used causal modeling proceeded in the following way: they used a "constraint-based method" to establish a local causal graph in the neighborhood of the target, using conditional independence tests. They then extracted a feature subset from this neighborhood and used it to build a predictive model. The predictive models used belong to the family of regularized discriminant classifiers and include L1-penalized logistic regression, ridge regression, and Support Vector Machines (SVM). Descriptions of these methods are found *e.g.,* in (Hastie et al., 2000). The feature subsets extracted from the local causal graph differ according to the test set distribution. For unmanipulated test sets (sets numbered 0), the Markov blanket of the target is chosen, including only direct causes (parents), direct effects (children), and spouses. For test sets drawn from post-manipulation distributions (numbered 1 and 2), two cases arise: if the identity of the manipulated features is known to the participants, the feature subset selected is a restriction of the Markov blanket to parents (direct causes), unmanipulated children (direct effects), and parents of at least one unmanipulated child (spouses). This is the case for REGED1 and MARTI1. If the identify of the manipulated features is unknown to the participants, the feature subset selected is limited to direct causes. The techniques used to learn the local causal graph from training data are all derived from the work of Aliferis and Tsamardinos and their collaborators (Aliferis et al., 2003a; Tsamardinos and Aliferis, 2003; Aliferis et al., 2003b). Gavin Cawley (Cawley, 2008) used directly the "Causal explorer"

package provided by the authors (Aliferis et al., 2003b). Laura E. Brown and Ioannis Tsamardinos (L.E.B. & Y.T.) (Brown and Tsamardinos, 2008) improved on their own algorithms by adding methods for overcoming several simplifying assumptions like "faithfulness" and "causal sufficiency". They proposed to address the problem of faithfulness by using a method for efficiently selecting products of features, which may be relevant to predicting the target, using non-linear SVMs, and proposed to address the problem of violations of causal sufficiency and hidden confounders by examining so-called "Y structures". Jianxin Yin and Zhi Geng's group (Yin et al., 2008) also introduced elements of novelty by proceeding in several steps: (1) removing features which are surely independent, (2) looking for parents, children, and descendants of the target and identify all V-structures in the neighborhood of the target, (3) orienting as many edges as possible, (4) selecting a suitable restriction of the Markov blanket (depending on the test set distribution, as explained above), (5) using L1-penalized logistic regression to assess the goodness of causal discover and eventually removing remaining redundant of useless features.

Markov blanket discovery: Discovering the Markov blanket is a by-product of causal discovery algorithms and can also sometimes be thought of as a sub-task. If known exactly, the Markov blanket is a sufficient set of features to obtain best prediction results if the test data are not manipulated. As explained in the previous paragraph, to remain optimal, this feature set must be restricted in the case of manipulated test data to parents (direct causes), unmanipulated children, and parents of unmanipulated children, or to only direct causes (depending on whether the manipulations are known or not). Hence, using the Markov blanket of the natural distribution for all test sets, including those drawn from post-manipulated distributions, is in principle sub-optimal. However, several participants adopted this strategy. One noteworthy contribution is that of Mehreen Saeed (Saeed, 2008), who proposed a new fast method to extract the Markov blanket using Dirichlet mixtures.

Feature selection: There have been a wide variety of feature selection methods, which have proved to work well in practice in past challenges (Guyon et al., 2006a). They do not have any theoretical justification of optimality for the causal discovery problem, except that in some cases it can be proved that they approximate the Markov blanket (Nilsson et al., 2007). Several participants used feature selection methods, disregarding the causal discovery problem, and obtained surprisingly good results. See our analysis in Section 6. The methods employed belong to the family of "filters", "wrappers" or "embedded methods". Vladimir Nikulin (Nikulin, 2008) used a "wrapper" approach, which can be combined with any learning machine, treated as a "black box". The method consists in sampling feature sets at random and evaluating them by cross-validation according their predictive power using any given learning machine. The features appearing most often in the most predictive subsets are then retained. Yin-Wen Chang and Chih-Jen Lin (Chang and Lin, 2008), Gavin Cawley (Cawley, 2008), and Jianxin Yin and Zhi Geng's group (Yin et al., 2008) used embedded feature selection methods relying on the fact that, in regularized linear discriminant classifiers, the features corresponding to weights of small magnitude can be eliminated without performance degradation. Such methods are generalizable to non-linear kernel methods via the use of scaling factors. They include RFE-SVM (Guyon et al., 2002) and L1-penalized logistic or ridge regression (Tibshirani, 1994; Bi et al., 2003). Marc Boullé (M.B.) used a univariate filter method making assumptions of independence between variables.

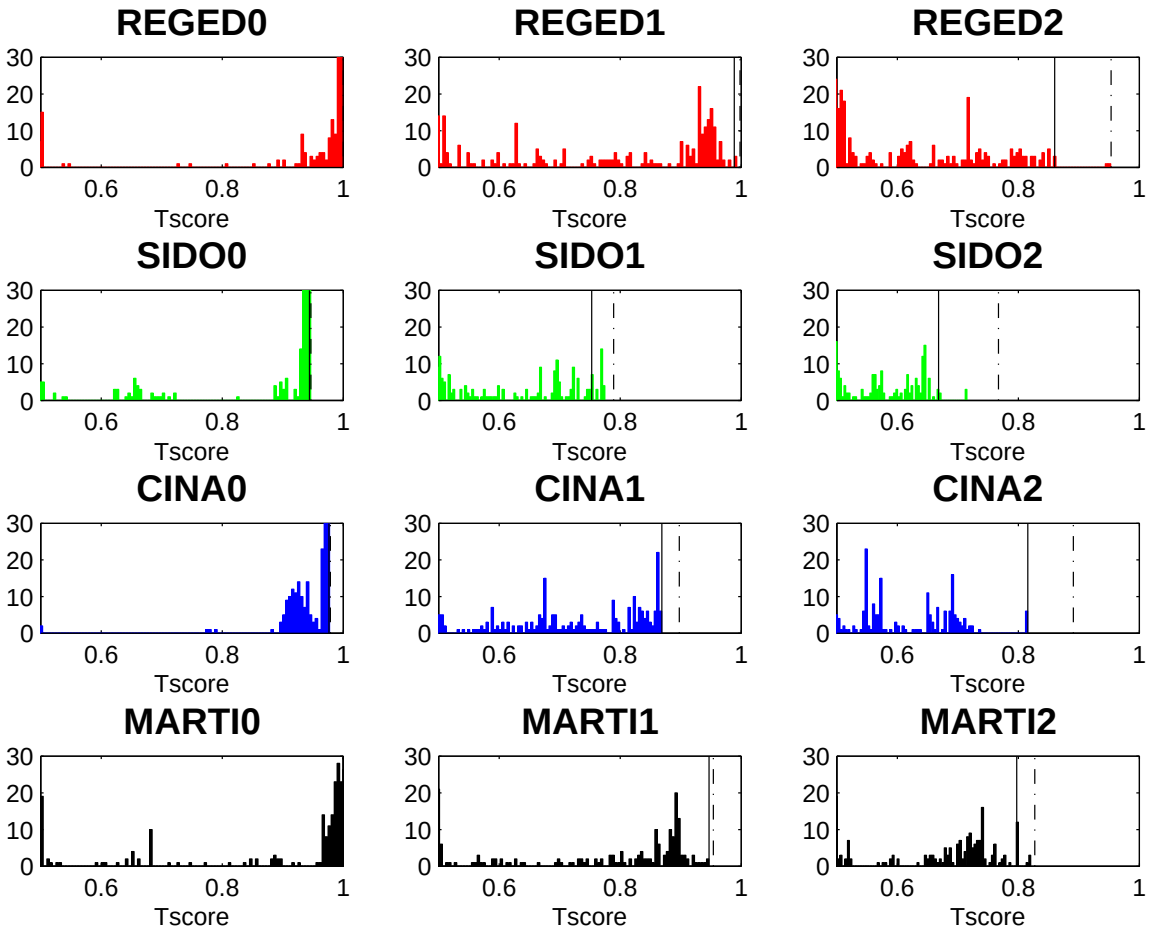


Figure 3: **Performance histograms.** We show histograms of Tscore for all entries made during the challenge. The vertical solid line indicates the best ranked entry (*i.e.,* best among the last complete entries of all participants). The dashed line indicates the overall best, including Reference entries, utilizing the knowledge of causal relationships not available to participants.

5.4 Analysis by dataset

We show in Figure 3 histograms of the performances of the participants for all the entries made during the challenge. We also indicate on the graphs the positions of the best entry counting towards the final participant ranking; *i.e.,* their last complete entry, and the very best entry (among all entries including Reference entries made by the organizers.) As can be seen, the distributions are very different across tasks and test set types. In what follows, we discuss specific results.

For the test sets numbered 0, the best entries closely match the best Reference entries made by the organizers, who used knowledge of feature relevance not available to the competitors (such Reference entries used the Markov blanket of the target variable in the post-manipulation distribution as feature set and a SVM classifier). This is encouraging and shows the maturity of feature selection techniques, whether they are based or not on the extraction of the Markov blanket. For two datasets (REGED0 and CINA0), the univariate method of Marc Boullé (M.B.), which is based on the *naïve Bayes assumption* (independence between features) was best. This method had already shown its strength in previous challenges on the Adult database based on census data, from which CINA is derived. Interestingly, we know by construction of the REGED dataset that the naïve Bayes assumption does not hold, and yet good performance was obtained. This result is a nice illustration of the bias *vs.* variance tradeoff, which can lead biased models to yield superior prediction accuracy when training data are scarce or noisy. In this case, the multivariate methods of Yin-Wen Chang and Chih-Jen Lin (Chang and Lin, 2008) for REGED0 and Vladimir Nikulin (Nikulin, 2008) for CINA0 have results which are not significantly different from the univariate method of Marc Boullé.[14] For SIDO0, the best results achieved by Jianxin Yin and Zhi Geng's group (Yin et al., 2008) are not significantly different from the results of Gavin Cawley (Cawley, 2008), using *no feature selection.* Generally, regularized classifiers have proved to be insensitive to the presence of irrelevant features, and this results confirms observations made in past challenges (Guyon et al., 2006a,b, 2008b). The best result for MARTI0 is also obtained by Gavin Cawley. His good performance can probably be partially attributed to the sophistication of his preprocessing, which allowed him to remove the correlated noise. In a post-challenge comparison he conducted between a Markov blanket-based feature selection and BLogReg, an embedded method of feature selection based on regularization, both methods performed well and the results were not statistically significantly different, and interestingly the BLogReg method yielded fewer features than the Markov blanket-based method.

For the test sets numbered 1 and 2, the distribution of test data differed from the training data. There is still on several datasets a large difference between the results of the best entrants and the best achievable result estimated by the organizers, using the knowledge of the true causal relationships. Sets 2 were more difficult than sets 1, for various reasons, having to do with the type of manipulations performed. Rather surprisingly, for test sets 1, non-causal methods yielded again very good results. Marc Boullé (M.B.) obtained the best performance on REGED1, with his univariate method making independence assumptions between features and involving no causal discovery. His feature set of 122/999 features

14. In the rest of this analysis, "not significantly different" means within one sigma, using our approximate error bar of Equation 1.

does not contain variables which are not predictive (*i.e.,* not connected to the target in the causal graph of the post-manipulation distribution), but in general there are very few such variables. The best ranked competitor on REGED1, Marius Popescu, uses a particularly compact subset of 11 features, obtained with a causal discovery method combining HITON-MB (Aliferis et al., 2003a), some heuristics to orient edges, and the elimination of manipulated children and spouses whose children are all manipulated (see the Fact Sheet for details). According to pairwise comparisons, the next best result is obtained by the causal method of Laura E. Brown and Ioannis Tsamardinos (L.E.B. & Y.T.) (Brown and Tsamardinos, 2008) with only 9 features. Their feature set does not coincide exactly with the Markov blanket of the post-manipulation distribution (which includes 14 features), but it contains no irrelevant feature. For SIDO1, the best performance was obtained with all or nearly all features by Jianming Jin (Yin et al., 2008), Yin-Wen Chang (Chang and Lin, 2008), and Gavin Cawley (Cawley, 2008). Hence, even when manipulations are performed, feature selection is so hard on this dataset that one is better off not doing any feature selection. The best performing causal discovery method on this dataset is that of Jianxin Yin and Zhi Geng's group (Yin et al., 2008), but their performance is significantly lower than that obtained with no feature selection (Tscore 0.70 instead of 0.75, with an error bar of 0.01). For CINA1 and MARTI1, Vladimir Nikulin (Nikulin, 2008) obtains the best performances with a feature selection method in pairwise comparisons (even though Gavin Cawley comes ahead in Table 2). He uses 30/132 and 400/1024 features, respectively. His fraction of irrelevant features in CINA1 is no better than the original proportion on the entire feature set. Jianxin Yin and Zhi Geng's group (Yin et al., 2008) are second best on those two datasets, with performances which are not statistically significantly different. Their causal discovery method yields fewer features (24/132 and 11/1024) and a smaller fraction of irrelevant features. The next best entries include both causal and non-causal methods. In conclusion, neither causal discovery methods nor feature selection methods seem to come ahead on test sets 1. This result will be further discussed in Section 6.

For test sets 2, making good predictions without causal modeling was expected to be significantly harder. Yet Jianxin Yin and Zhi Geng's group (Yin et al., 2008) are the only ones using causal discovery performing consistently well on sets 2. They are first or second in pairwise comparisons for REGED2, SIDO2, and MARTI2. For REGED2, Laura E. Brown & Ioannis Tsamardinos (L.E.B. & Y.T.) obtained the best performance with a causal discovery method. For SIDO2, E. Mwebaze and J. Quinn perform best according to pairwise comparisons, also using a causal discovery method. But, for MARTI2, none of the other top-ranking entries (in pairwise comparisons) include causal discovery methods, even though there is a very significant correlation between Fscore and Tscore (0.88). We will discuss the case of MARTI2 in more detail in Section 6. On CINA2 all causal discovery methods perform poorly, except that of Florin Popescu. However, since he submitted results only on CINA, it is difficult to say whether he was lucky or has a causal discovery method that is competitive on this problem. The other methods, which did well on CINA according to pairwise comparisons are those of Marc Boullé (M.B.) (naïve Bayes) and Vladimir Nikulin (feature selection with the random subset method). When selecting the best results in nested subsets of features, Yin-Wen Chang obtained significantly better results than anyone else with her SVM feature ranking. Her best feature set included only 4 features, which are all "real" variables. The case of CINA will be further discussed in Section 6.

6. Discussion

Several algorithms have demonstrated effectiveness of discovering causal relationships, as indicated by the Fscore, hence this challenge contributed to demonstrating that causal discovery from observational data is not an impossible task, albeit a very hard one. Yet the performance of causal models on tasks that were purposely designed to demonstrate their effectiveness is somewhat disappointing. It can be argued that causal discovery is a relatively new domain of research, which has not yet reached the maturity of some of the more mainstream machine learning techniques that were applied with success to the challenge. In particular, the use of causal discovery software made freely available may not be straightforward to use appropriately for people new to the field. However, it seems plausible that other factors are at play. In this section we analyze the results of the challenge in a critical manner and invite researchers to further investigate the open problems.

6.1 Correlation between causation and prediction in an interventional setting

One of our main motivations in organizing this challenge was to investigate the extent to which causal modeling is useful for making predictions in an "interventional setting" (a setting in which the test set is distributed differently from the training set as a result of the intervention of an external agent). Hence, in our analysis, we tried to quantify the correlation between "causation" (the accuracy of the causal modeling around the target variable) and "prediction" (the accuracy of the target variable predictions on test data). The former is captured by the Fscore and the latter by the Tscore. After modifying the Fscore, and averaging over all datasets and test set versions, we obtain a significant correlation between Fscore and Tscore (pvalue 2.10^{-19}). But, for individual tasks, there is a lot of variability. In past challenges (Guyon et al., 2006a,b, 2008b), it was already observed that feature selection does not necessarily improve prediction accuracy when training and test data are drawn from the same distribution. This is due to the fact that state-of-the-art regularized classifiers such as SVMs, ridge regression, Random Forests (RF) and ensembles of neural networks, effectively overcome the curse of dimensionality without requiring a dimensionality reduction performed as preprocessing. In fact, feature selection is sometimes more harmful than useful in this case. For example, the best result on SIDO0 is obtained with no feature selection (in spite of the presence of irrelevant artificial variables or "probes"). More surprisingly, for test sets 1 and 2, although there is a significant correlation between Fscore and Tscore (on average over all tasks), we observe that feature selection methods based on causal discovery methods rarely outperforms feature selection methods ignoring causal relationships.

In a recent analysis paper (Tillman and Spirtes, 2008), the authors investigate the total contribution to prediction error made when non-causal methods use incorrect predictors for a manipulated distribution and when causal methods use incorrect or biased parametric constraints. They give theoretical conditions for manipulations where causal methods for prediction should have no advantage over non-causal methods and for manipulations where causal methods should produce considerably fewer errors. Briefly, the post-manipulation distribution $P(target|predictors)$ is identical to the natural distribution $P(target|predictors)$ only under special conditions, including that there is no manipulated direct effect of the target in the predictor set. The most difficult cases for non-causal methods arise (1) when

all variables are manipulated or (2) when the non-manipulated variables (other than the target) are sampled before the effect of the manipulations have propagated.

Following this line of reasoning, we can partially explain why non causal methods performed so well by examining our challenge design. We sampled the variables after the effect of the manipulations propagated, because of the nature of our applications. Had we sampled them before the effect of the manipulations propagated, we would have made the task harder for non causal methods. Far fewer variables would have been predictive of the target, and, in particular, no consequence of the target would have been predictive. However, this limitation of our design was partially compensated by manipulating a large number of variables, including many direct effects of the target. Consequently, non causal methods incurred a larger false positive rate than causal methods for test sets 1 and 2, because many features relevant in the natural distribution were irrelevant in the post-manipulation distribution.

In the next section, we propose another explanation, which sheds light on the difficulty of improving performance with any kind of feature selection, causal or not.

6.2 Omitting good features may be more detrimental than including bad ones

We provide a qualitative explanation of why selecting a relatively large fraction of "irrelevant" features (including features relevant in training data and irrelevant in test data) might not penalize as much predictions as omitting key "relevant" features. The idea of our argument is that, in a predictive model, relevant variables tend to act in the same direction (to build the predictive signal) while, in the large sample limit, irrelevant variables contribute signals which average out to zero.

The results of our calculations provided in Appendix A are summarized in Table 3. We see that for test sets 0 the contribution of the irrelevant features can be very small compared to that of relevant features. To evaluate the number of irrelevant features one can "afford" for a given number of relevant features, we use the crossing point where the contribution of both type of features is equal. We obtain, for test sets 0, $n_b = m\ n_g^2$, for test sets 1, $n_b = n_g^2$, and for test sets 2, $n_b = n_g$. Plugging in some numbers, if there are of the order of $n_g = 30$ relevant features and $m = 500$ training examples, one can afford for test sets 0 of the order of $n_b = m \cdot n_g^2 = 500 \times 900 = 450,000$ irrelevant features. No wonder feature selection is not all that important in that case. For test sets 1, it is not that critical either to filter out irrelevant features, even if they are relevant in training data and manipulated in test data. Plugging in some numbers, if there are of the order of $n_g = 30$ relevant features, we can afford of the order of $n_b = n_g^2 = 900$ irrelevant features. This largely explains that feature selection is more needed on sets 1 than on sets 0, but simple feature selection does as well as causal feature selection. Finally, only in the worst case scenario of adversarial manipulations (test sets 2), can we only afford a number of "bad" features of the same order of magnitude of the number of "good" features.

The properties of irrelevant variables, on the basis of which we conclude that omitting relevant variable might more severely impair performance than including irrelevant variables, are obviously distribution dependent, and a case-per-case analysis would be needed to make a more quantitative assessment. We also need to caution against extrapolating our qualitative explanations and concluding that there is no benefit to performing causal dis-

Table 3: **Noise introduced by irrelevant features.** We computed for a simple univariate predictive model, the influence of relevant and irrelevant features. Both features and target are binary, and it is assumed that all relevant features correlate perfectly with the target and all irrelevant features are randomly drawn. With 98% confidence, the magnitude of the feature weights are lower than the value w quoted in the table and the total contribution $\sum_i w_i x_i$ is lower than the v quoted. n_g is the number of "good" (relevant) features and n_b is the number of "bad" (irrelevant) features, and m is the number of training examples.

Test set	Type	w relevant	w irrelevant	v relevant	v irrelevant
Set 0	unmanipulated	1	$1/\sqrt{m}$	n_g	$\sqrt{n_b/m}$
Set 1	manipulated	1	1	n_g	$\sqrt{n_b}$
Set 2	manipulated	1	1	n_g	n_b

covery because of the relative insensitivity of certain regularized classifiers to the presence of irrelevant features (those who have parameters acting as feature scaling factors). These conclusions apply only to the particular tasks of the challenge and modifying the tasks may yield different conclusions. For example, the problem of finding which variables are the best targets of action to obtain a desired response requires a causal model. Also, we could have made things more difficult to non-causal models by sampling before the effect of the manipulations propagate to the non-target variables, thus making all consequences of the target variable non-predictive. However, this would not have affected the performances for test sets 2 in which all effect or all probes are manipulated.

6.3 Insignificant dependencies and spurious dependencies

We are left with explaining why for CINA2 and MARTI2 causal discovery methods do not perform as well as expected.

For CINA, we attribute the problem to the variable coding, which diluted information and led to many insignificant dependencies. By examining the features selected by regular feature selection algorithms and by causal discovery algorithms, we noticed that they were rather different. Feature selection algorithms select features that are individually very predictive, but not part of the Markov blanket. This may be due to the coding of categorical variables that we used: categorical variables taking c values were replaced by c binary variables, implementing a complete disjunctive code 10...0, 01...0, etc. So for instance, "number of years of education", which may be an ancestor variable of "profession", is individually more predictive than any of the individual professions: "clerical", "managerial", etc. Verifications of this explanation are under way by examining the causal graphs inferred by the top-ranking participants and the results will be published as part of the analysis of the second causality challenge (Guyon et al., 2008a).

For MARTI2, the correlated noise was considerably difficult for causal discovery algorithms, which did not perform well unless the noise was efficiently filtered. This is confirmed by the fact that causal discovery methods did well on REGED, a noiseless version of MARTI.

6.4 Is the Markov blanket truly optimal?

The above consideration on the relative insensitivity of predictive modeling to the presence of "irrelevant" variables may not alone explain the good performances of feature selection methods. It is possible that restricting the feature set to the Markov blanket of the test set distribution hampered performances. This strategy was adopted by all the participants performing causal discovery. If the causal paths to the target are not interrupted by the manipulations, adding some predictive non-MB variables (like the light green nodes in Figure 1) may help improving performances when a biased classifier is used, *e.g.,* if the MB is non-linearly related to the target and a linear classifier is used (Tsamardinos and Aliferis, 2003). Furthermore, the MB may include errors when estimated from a finite training set. We noted in the above calculations that it is far more detrimental to omit relevant features than include irrelevant features. Hence, subsets of larger size than the estimated MB are likely to give better predictions. The strategy adopted by the participants performing causal discovery was also sub-optimal in another respect: they selected a subset S of features that should be predictive of the target Y in the post-manipulation distribution, then they trained a "regular" learning machine to estimate directly $P(Y|S)$ with training data from the natural distribution. In some cases, this is *not* equivalent to estimating $P(Y|S)$ in the post-manipulation distribution by using a causal model. Cases of that sort arise when one manipulates children of the target, which have *unmanipulated* children of the target as descendants. For instance, in the LUCAS example of Figure 1, if we had manipulated the variable "coughing", but not the variable "fatigue", both "fatigue" and "coughing" would still be in the Markov Blanket of the post-manipulation distribution ("coughing" would now be a "spouse" of the target), but the direct connection between the target and "coughing" would be broken. Hence the contribution of "coughing" to $P(Y|S)$ would be over-estimated if $P(Y|S)$ was estimated by a statistical learning machine trained from the natural distribution, because this would include the direct effect of the target on "coughing". In contrast, a causal model taking into account the manipulations would factor out such direct effect when estimating $P(Y|S)$.

6.5 Lessons learned for future challenges

We end this discussion with some comments on the challenge protocol. First, as noted before, selecting the best Tscore in nested subsets of features introduced bias in the results and we do not recommend using this paradigm in future challenges of this kind. It is necessary to ask the participants to provide single predictions, or make pairwise comparisons of performance at equal number of features. Second, in our setup, the target variable was never manipulated. This makes sense for problems in which we are seeking to discover causes of a given outcome in order to influence it. For example, in epidemiology, we want to find risk factors of lung cancer such as smoking. But there are problems in which a target variable is manipulated and the goal is to monitor the effects of the manipulation. For example, the disappearance of symptoms can help monitoring the effect of a drug on a disease. Third, in our setup, we perform manipulations and wait before we sample data, until the effects of the manipulations have propagated through the system. In some cases, it makes more sense to sample data before the manipulations are performed and ask the question: what if we did these manipulations to given variables? Fourth, in our setup, we

asked the participants to make predictions of a target variable under manipulations of other variables. Emphasis was placed on prediction rather than on variable selection. Another question would be to find those variables which should be manipulated to produce a given desired effect, *i.e.,* a given change in the target value. Finally, we posed a problem in which causal models had to be inferred solely from observational data. In many cases, it is costly but feasible to include manipulated data as part of training.

7. Conclusions

The first causality challenge we have organized allowed many researchers both from the causal discovery community and the machine learning community to try their algorithms on sizeable tasks of real practical interest. It achieved a number of goals that we had set: familiarizing many new researchers and practitioners with causal discovery problems and existing tools to address them, pointing out the limitations of current methods on some particular difficulties, and fostering the development of new algorithms.

The setting of the challenge purposely resembled a classical machine learning competition, with a training set and a test set, with omitted labels, to encourage the participation of data mining and machine learning researchers. The goal was to make optimum predictions on test data, as measured by a Tscore (the area under the ROC curve on test data). Each task had three test sets, with increasing levels of difficulty. The first one was identically distributed as the training set. The two other test sets simulated manipulations by external agents, and thus were not distributed like the training set. In this way we illustrated the relationships between causation and prediction under manipulations and investigated whether causal models using "causally relevant" features would perform better than regular statistical models on manipulated test sets. We proposed a simple score to evaluate the causal relevance of the subset of features selected, called Fscore. Several algorithms have demonstrated effectiveness of discovering causal relationships, as indicated by a large Fscore. On average over all datasets and tasks, the Fscore correlates significantly with the Tscore, confirming the link between causation and prediction. As anticipated, non-causal feature selection methods are doing well on the first type of datasets (training and test data identically distributed): the bulk of them is close to optimal, so if you chose one method at random, you would do well. However, for the other two types of datasets (test data manipulated) the distribution of results is about uniform: if you chose one method at random, you would probably do poorly. In addition, there is room for improvement to reach optimality. Thus, non-causal feature selection methods are inappropriate for these tasks, despite the fact that some of them are top ranked and causal feature selection methods are still not mature and robust enough to significantly outperform non-causal feature selection in the range of tasks of the competition.

The results indicate that informative causal prediction from observational data is possible, although it remains challenging. This points to the need for further research and benchmarks. This challenge investigated an important problem in causal modeling, but there remain many other causal modeling and discovery issues to be explored. Future work includes organizing challenges on a broader range of causal questions.

Acknowledgments

The organization of this challenge was a team effort to which many contributed. We are particularly indebted to Olivier Guyon (MisterP.net) who implemented the back-end of the web site. The front-end is derived from the design of Steve Gunn (University of Southampton), formerly used for the NIPS 2003 feature selection challenge. The kind support of Joachim Buhmann (ETH Zurich), who hosted the competition, is gratefully acknowledged. We are thankful to Thomas Fuchs (ETH Zurich) for administering the computer resources. The beta-testers, Gideon Dror (Academic College, Tel-Aviv-Yaffo, Israel), Amir Saffari (Graz University of Technology, Austria) and Marc Boullé (France Telecom, Lanion, France) performed indispensable work to ensure the platform was working properly, give us feed-back on the protocols and creating "Reference" entries. Alexander Borisov was first person to enter the challenge and he provided us with invaluable feed-back, which helped building the FAQ, which was an important contribution. We also are thankful to Gavin Cawley (University of East Anglia, UK), Joaquin Quiñonero Candale (Microsoft Research, UK), Richard Scheines (Carnegie Mellon University, Pennsylvania), and Lambert Schomaker (University of Groningen, The Netherlands) for helpful advice. We are very grateful to the institutions who originally provided the data: the DTP AIDS Antiviral Screen program of the National Cancer Institute (NCI); the census bureau and Ronny Kohavi and Barry Becker who extracted the data. Hans Bitter (Roche, Palo Alto, California) and Joerg Wichard (Institute of Molecular Pharmacology, Berlin, Germany) who provided features for SIDO are gratefully acknowledged. This project was supported by the Pascal network of excellence funded by the European Commission and by the U.S. National Science Foundation under Grant N0. ECCS-0725746. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the National Science Foundation. Additional support was provided by the U.S. National Institute of Health under grant 2R56LM007948-04A1. The WCCI2008 conference, Health Discovery Corporation, Microsoft, and Unipen provided additional support for the monetary prizes granted to the winners. The comments of Ioannis Tsamardinos were very helpful to improve the manuscript, and we thank him for his effort.

References

C. F. Aliferis, I. Tsamardinos, and A. Statnikov. HITON, a novel Markov blanket algorithm for optimal variable selection. In *2003 American Medical Informatics Association (AMIA) Annual Symposium*, pages 21–25, 2003a.

C. F. Aliferis, I. Tsamardinos, A. Statnikov, and L.E. Brown. Causal explorer: A probabilistic network learning toolkit for biomedical discovery. In *2003 International Conference on Mathematics and Engineering Techniques in Medicine and Biological Sciences (METMBS)*, Las Vegas, Nevada, USA, June 23-26 2003b. CSREA Press.

J. Bi, K. Bennett, M. Embrechts, C. Breneman, and M. Song. Dimensionality reduction via sparse support vector machines. *JMLR*, 3:1229–1243, 2003.

M. Boullé. Compression-based averaging of selective naive bayes classifiers. *JMLR*, 8: 1659–1685, July 2007a.

M. Boullé. Report on preliminary experiments with data grid models in the agnostic learning vs. prior knowledge challenge. In *IEEE/INNS conference IJCNN 2007*, Orlando, Florida, August 12-17 2007b.

L. E. Brown and I. Tsamardinos. A strategy for making predictions under manipulation. In *JMLR W&CP*, volume 3, pages 35–52, WCCI2008 workshop on causality, Hong Kong, June 3-4 2008.

G. Cawley. Causal and non-causal feature selection for ridge regression. In *JMLR W&CP*, volume 3, WCCI2008 workshop on causality, Hong Kong, June 3-4 2008.

Y.W. Chang and C.J. Lin. Feature ranking using linear svm. In *JMLR W&CP*, volume 3, pages 53–64, WCCI2008 workshop on causality, Hong Kong, June 3-4 2008.

I. Guyon, C. Aliferis, G. Cooper, A. Elisseeff, J.-P. Pellet, P. Spirtes, and A. Statnikov. Design and analysis of the causality pot-luck challenge. In *JMLR W&CP*, volume 5: NIPS 2008 causality workshop, to appear, Whistler, Canada, December 12 2008a.

I. Guyon, C. Aliferis, and A. Elisseeff. *Causal Feature Selection*, pages 63–82. Chapman and Hall/CRC Press. Longer TR: http://clopinet.com/isabelle/Papers/causalFS.pdf, 2007.

I. Guyon, S. Gunn, M. Nikravesh, and L. Zadeh, Editors. *Feature Extraction, Foundations and Applications*. Studies in Fuzziness and Soft Computing. With data, results and sample code for the NIPS 2003 feature selection challenge. Physica-Verlag, Springer, 2006a.

I. Guyon, A. Saffari, G. Dror, and J. Buhmann. Performance prediction challenge. In *IEEE/INNS conference IJCNN 2006*, Vancouver, Canada, July 16-21 2006b.

I. Guyon, A. Saffari, G. Dror, and G. Cawley. Analysis of the IJCNN 2007 agnostic learning vs. prior knowledge challenge. In *Neural Networks*, volume 21, pages 544–550, Orlando, Florida, March 2008b.

I. Guyon, J. Weston, S. Barnhill, and V. Vapnik. Gene selection for cancer classification using support vector machines. *Machine Learning*, 46(1-3):389–422, 2002.

I. Guyon et al. Datasets of the causation and prediction challenge. Technical Report, 2008.

T. Hastie, R. Tibshirani, and J. Friedman. *The Elements of Statistical Learning, Data Mining, Inference and Prediction*. Springer Verlag, 2000.

R. E. Neapolitan. *Learning Bayesian Networks*. Prentice Hall series in Artificial Intelligence. Prentice Hall, 2003.

V. Nikulin. Random sets approach and its applications. In *JMLR W&CP*, volume 3, pages 65–76, WCCI2008 workshop on causality, Hong Kong, June 3-4 2008.

R. Nilsson, J. M. Peña, J. Björkegren, and J. Tegnér. Consistent feature selection for pattern recognition in polynomial time. *J. Mach. Learn. Res.*, 8:589–612, 2007. ISSN 1533-7928.

Judea Pearl. *Causality: models, reasoning and inference*. Cambridge University Press, March 2000.

J. Quiñonero Candela, A. Schwaighofer, and N. Lawrence. Learning when test and training inputs have different distributions, http://different.kyb.tuebingen.mpg.de/pages/home.php 2007.

M. Saeed. Bernoulli mixture models for markov blanket filtering and classification. In *JMLR W&CP*, volume 3, pages 77–91, WCCI2008 workshop on causality, Hong Kong, June 3-4 2008.

P. Spirtes, C. Glymour, and R. Scheines. *Causation, Prediction, and Search*. The MIT Press, Cambridge, Massachusetts, London, England, 2000.

R. Tibshirani. Regression selection and shrinkage via the lasso. Technical report, Stanford University, Palo Alto, CA, June 1994.

R. E. Tillman and P. Spirtes. When causality matters for prediction: Investigating the practical tradeoffs. In *JMLR W&CP*, volume 5: NIPS 2008 causality workshop, to appear, Whistler, Canada, December 12 2008.

I. Tsamardinos and C.F. Aliferis. Towards principled feature selection: Relevance, filters, and wrappers. In *Ninth International Workshop on Artificial Intelligence and Statistics*, Florida, USA, January 2003.

J. Yin, Y. Zhou, C. Wang, P. He, C. Zheng, and Z. Geng. Partial orientation and local structural learning of causal networks for prediction. In *JMLR W&CP*, volume 3, pages 93–104, WCCI2008 workshop on causality, Hong Kong, June 3-4 2008.

Appendix A. Influence of irrelevant variables

We made an argument that adding irrelevant variables to the predictive feature set might not be as detrimental as omitting good ones. We base our qualitative analysis on a simple model, assuming that all variables including the target are binary (taking values ± 1) and that we use a linear predictive model

$$v = \sum_i w_i x_i,$$

in which the weights are trained with "Hebb's rule"

$$w_i = (1/m) \sum_k x_i^k y^k,$$

where the index k runs over all training examples, and m is the number of training examples. We further assume that the features are either perfectly relevant (identical to y or $-y$) or perfectly irrelevant (random). We wish to compute the relative contribution of relevant and irrelevant features to v in various cases to give insight into the number of irrelevant features, which can be afforded, relatively to the number of good features selected. In all cases, the magnitude (absolute value) of the weight of relevant features are:

$$w_{relevant} = 1.$$

Hence, the overall contribution of relevant features is the number of relevant or "good" features:

$$v_{relevant} = n_g.$$

For irrelevant features, we first examine the case where training and test data are identically distributed (case 0). If the irrelevant features are drawn randomly with equal probability $p = 0.5$, then the expected value of the magnitude of the weights of irrelevant features is 0. The standard deviation of the mean of $x_i^k y^k$ is $\sqrt{p(1-p)/m} = 0.5/\sqrt{m}$. To simplify our calculation, we use 98% confidence intervals, which roughly correspond to 2 sigma error bars by approximating the Binomial distribution with the Normal law. Hence, with 98% confidence, the magnitude of the weights of irrelevant features is less than

$$w^0_{irrelevant} = 1/\sqrt{m}.$$

We therefore verify that, for this model, the contribution of the irrelevant features vanishes to zero in the large sample limit. Similarly, the test set values of x_i are drawn randomly with equal probability $p = 0.5$. Hence, the total contribution has mean 0, and standard deviation bounded by $w\sqrt{n_b\ p(1-p)} = 0.5\ w\sqrt{n_b}$, where n_b is the number of irrelevant or "bad" features and w is our bound on the weight magnitude: $1/\sqrt{m}$. If we again choose a 98% confidence, we obtain a bound on the total contribution of the irrelevant variables of

$$v^0_{irrelevant} = \sqrt{n_b/m}.$$

In contrast, for test sets 1 and 2, in the worst case scenario, a feature perfectly relevant with respect to the training data distribution and perfectly irrelevant in the post-manipulated distribution will receive a weight of magnitude

$$w^1_{irrelevant} = w^2_{irrelevant} = 1.$$

In the scenario of test sets 1, values for such manipulated features are drawn randomly with equal probability $p = 0.5$. Following a calculation previously done, the standard deviation is bounded by $w\sqrt{n_b\ p(1-p)}$, but this time $w = 1$! The resulting bound on the total contribution of the "bad" features is

$$v^1_{irrelevant} = \sqrt{n_b},$$

with at least 98% confidence, because we assumed a worst-case scenario. For test sets 2, adversarial values may be given to the manipulated features, *i.e.,* opposite values than those expected from the training data distribution. So, in the worse case, the total contribution of the bad features is

$$v^2_{irrelevant} = n_b.$$

Appendix B:

Verification of Challenge Results

The rules of the challenge prohibited the use of testing data for feature selection and building of the classifier model. However, all testing data with the exception of the response variable was available to the challenge participants. That is why we decided to verify several submissions from the challenge by studying and executing source codes of the participants on our computers. While doing this verification we paid close attention to ease of reproduction of the challenge results and involved computational resources. Such information will be very useful to practitioners who may decide to apply such algorithms to other datasets.

We have selected 6 challenge participants that provided us with software and code for verification: *Gavin Cawley*, *Yin−Wen Chang*, *J. Yin & Z. Geng Gr.*, *L.E.B & Y.T.*, *Vladimir Nikulin*, and *Mehreen Saeed*. The base verification dataset was selected to be REGED due to its empirical difficulty in the challenge and requirement for causal feature selection. However, *Vladimir Nikulin* provided source codes only for CINA dataset and *Mehreen Saeed* provided codes only for CINA and SIDO datasets. Thus, we decided to use CINA dataset for verification of these two participants. Out of six selected participants, three (*J. Yin & Z. Geng Gr.*, *L.E.B & Y.T.*, *Vladimir Nikulin*) used algorithms for selection of a single feature set and the remaining participants (*Gavin Cawley*, *Yin−Wen Chang*, *Mehreen Saeed*) used techniques to selected nested subsets of features.

The verification protocol consisted of two major steps: (i) manual reading of the source code to ensure that it does not employ testing data during feature selection and building of the classifier model and (ii) reproducing results of the challenge in a series of experiments. We considered the following experiments for versions 0, 1, and 2 of the datasets:

Experiment	*Description*
1	Exact reproduction of the challenge submission
2	Using reduced testing dataset with 500 samples (250 positives and 250 negatives, selected at random)
3	Using reduced testing dataset with 200 samples (100 positives and 100 negatives, selected at random)
4	Using reduced testing dataset with 500 samples, selected at random
5	Using reduced testing dataset with 200 samples, selected at random
6	Same as experiment 1 but with randomly permuted variables and samples
7	Same as experiment 2 but with randomly permuted variables and samples
8	Same as experiment 3 but with randomly permuted variables and samples
9	Same as experiment 4 but with randomly permuted variables and samples
10	Same as experiment 5 but with randomly permuted variables and samples

Experiment #1 was designed to verify that the automated code of a participant matched the challenge entry in terms of classification AUC. Experiments #2-5 were primarily used to confirm that the challenge participant did not use testing data to make inferences about the distribution.

Experiment #6 was intended to illustrate that the code both does not rely on hard-coded feature indices and is not sensitive to the ordering of variables. Finally, experiments #7-10 were seeking goals of both experiments #2-5 and #6.

First of all, our manual reading of the source codes confirmed that none of the selected challenge participants cheated by using testing data for training of the classifier or feature selection.

Figure 1 and Table 1 report classification AUC's for the above described experiments. The results for versions 0 of the datasets are not reported in Figure 1 because they have near-perfect reproducibility. In summary, the results of the selected challenge participants reproduced in all experiments.

The code submitted by the team *L.E.B. & Y.T.* includes the automation of a step that during the competition was performed manually. The authors declared that the automated step is as close as possible to the subjective method used during the competition. An implementation of the strategy proposed by the authors is now fully automated and produces reproducible and repeatable results.

In all experiments we used Xeon 2.8 GHz CPU's with 4 Gb RAM. For the REGED dataset, the slowest algorithm was the one by *Gavin Cawley* with t_{train} = ~20-30 hours and the fastest one was by *Yin−Wen Chang* with t_{train} = 1-2 minutes. For all other methods in the REGED dataset, $t_{train} \in$ [15 minutes, 2 hours]. For the CINA dataset, all methods have t_{train} < 1 hour. The testing time was negligible for all algorithms and datasets (t_{test} <2 minutes). All algorithms had relatively efficient implementations. The only exception is the code by *Gavin Cawley* that required >300 Mb for storage of the model for REGED datasets. Another inefficiency was observed in the code of *Yin−Wen Chang* that required ~4 Gb of RAM to apply a model to a testing set of 20,000 instances.

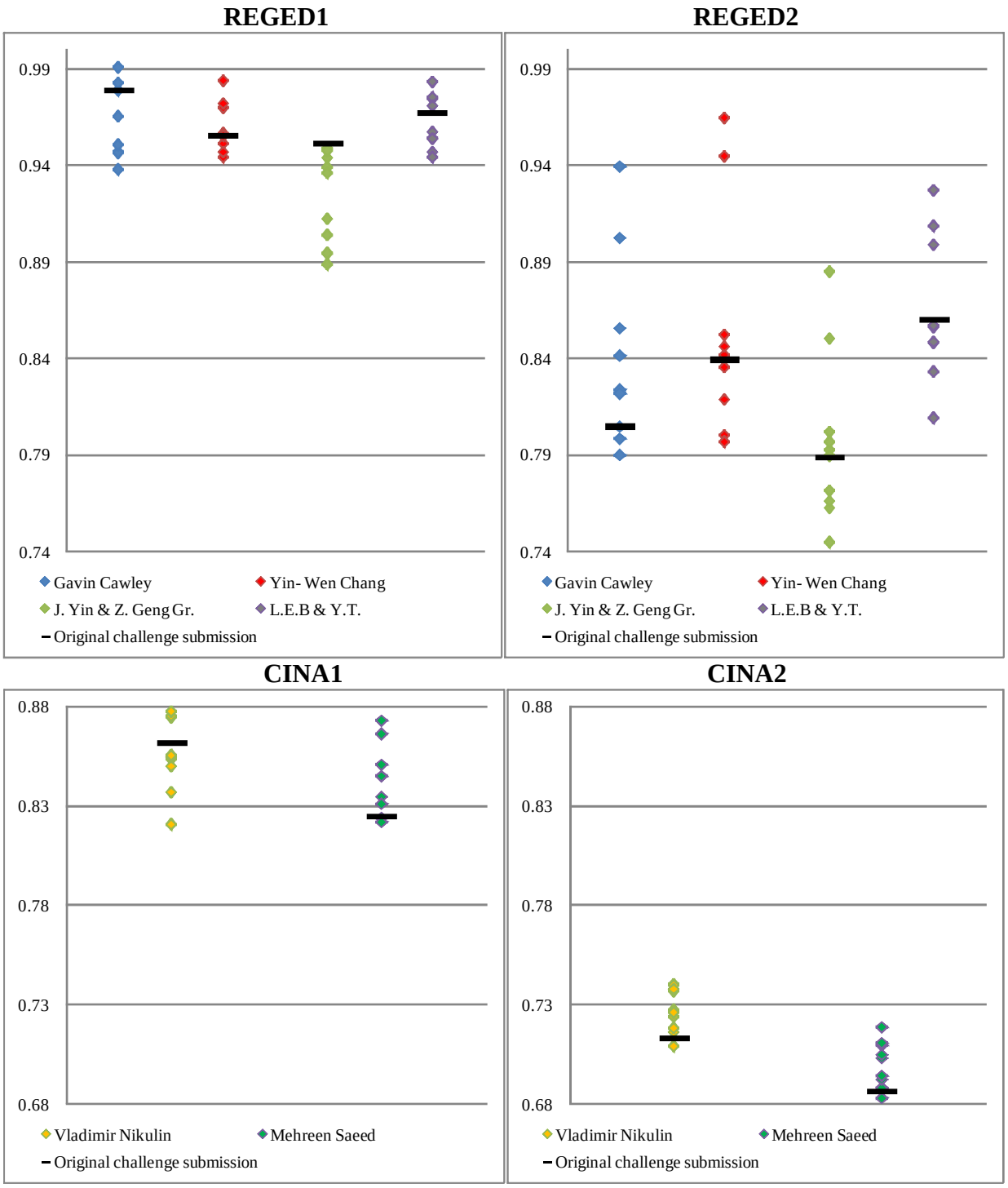


Figure 1: Testing set classification performance (measured by AUC) for 6 participants of the challenge. Each dot corresponds to results of an experiment.

	REGED0				CINA0	
Experiment	*Gavin Cawley*	*Yin-Wen Chang*	*J. Yin & Z. Geng Gr.*	*L.E.B & Y.T.*	*Vladimir Nikulin*	*Mehreen Saeed*
1	0.9997	0.9998	0.9998	0.9999	0.9770	0.9750
2	1.0000	0.9999	0.9998	0.9999	0.9812	0.9865
3	1.0000	1.0000	1.0000	1.0000	0.9512	0.9727
4	1.0000	0.9997	0.9997	0.9999	0.9752	0.9754
5	0.9983	0.9981	0.9997	0.9989	0.9805	0.9744
6	0.9998	0.9998	0.9998	0.9996	0.9760	0.9749
7	0.9998	1.0000	0.9999	0.9998	0.9741	0.9820
8	0.9997	0.9999	0.9996	0.9992	0.9622	0.9712
9	1.0000	1.0000	0.9999	0.9995	0.9709	0.9759
10	1.0000	1.0000	1.0000	1.0000	0.9682	0.9797
Challenge submission	**0.9997**	**0.9998**	**0.9997**	**0.9998**	**0.9764**	**0.9751**

	REGED1				CINA1	
Experiment	*Gavin Cawley*	*Yin-Wen Chang*	*J. Yin & Z. Geng Gr.*	*L.E.B & Y.T.*	*Vladimir Nikulin*	*Mehreen Saeed*
1	0.9787	0.9556	0.9442	0.9538	0.8549	0.8233
2	0.9789	0.9445	0.9392	0.9536	0.8532	0.8340
3	0.9825	0.9700	0.9490	0.9572	0.8742	0.8657
4	0.9907	0.9515	0.9478	0.9467	0.8366	0.8305
5	0.9905	0.9720	0.9362	0.9441	0.8206	0.8502
6	0.9469	0.9556	0.8943	0.9743	0.8542	0.8235
7	0.9463	0.9567	0.9125	0.9749	0.8552	0.8213
8	0.9506	0.9555	0.8888	0.9706	0.8746	0.8724
9	0.9378	0.9470	0.8947	0.9536	0.8497	0.8443
10	0.9653	0.9839	0.9042	0.9831	0.8769	0.8444
Challenge submission	**0.9787**	**0.9556**	**0.9517**	**0.9673**	**0.8617**	**0.8248**

	REGED2				CINA2	
Experiment	*Gavin Cawley*	*Yin-Wen Chang*	*J. Yin & Z. Geng Gr.*	*L.E.B & Y.T.*	*Vladimir Nikulin*	*Mehreen Saeed*
1	0.8045	0.8392	0.7926	0.8481	0.7159	0.6827
2	0.7984	0.8186	0.7626	0.8328	0.7401	0.7182
3	0.7897	0.8001	0.7660	0.8476	0.7367	0.7092
4	0.8218	0.7968	0.7448	0.8088	0.7238	0.7102
5	0.9025	0.9447	0.8850	0.9268	0.7088	0.6880
6	0.8237	0.8416	0.7896	0.8557	0.7180	0.6877
7	0.8218	0.8355	0.8019	0.8482	0.7271	0.7027
8	0.8413	0.8461	0.7968	0.8566	0.7258	0.6919
9	0.8555	0.8522	0.7715	0.8986	0.7398	0.7044
10	0.9395	0.9646	0.8503	0.9083	0.7377	0.6939
Challenge submission	**0.8045**	**0.8392**	**0.7885**	**0.8600**	**0.7132**	**0.6867**

Table 1: Testing set classification performance (measured by AUC) for 6 participants of the challenge.