Causal discovery of gene regulation with incomplete data

Ronja Foraita,

Leibniz Institute for Prevention Research and Epidemiology—BIPS, Bremen, Germany

Juliane Friemel,

Leibniz Institute for Prevention Research and Epidemiology—BIPS, Bremen, Germany, and University and University Hospital Zurich, Switzerland

Kathrin Günther,

Leibniz Institute for Prevention Research and Epidemiology—BIPS, Bremen, Germany

Thomas Behrens,

Ruhr University Bochum, Germany

Jörn Bullerdiek and Rolf Nimzyk

Center for Human Genetics, Bremen, Germany

and Wolfgang Ahrens and Vanessa Didelez

Leibniz Institute for Prevention Research and Epidemiology—BIPS, Bremen, and University of Bremen, Germany

Summary. Causal discovery algorithms aim to identify causal relations from observational data and have become a popular tool for analysing genetic regulatory systems. In this work, we applied causal discovery to obtain novel insights into the genetic regulation underlying head-and-neck squamous cell carcinoma. Some methodological challenges needed to be resolved first. The available data contained missing values, but most approaches to causal discovery require complete data. Hence, we propose a new procedure combining constraint-based causal discovery with multiple imputation. This is based on using Rubin's rules for pooling tests of conditional independence. A second challenge was that causal discovery relies on strong assumptions and can be rather unstable. To assess the robustness of our results, we supplemented our investigation with sensitivity analyses, including a non-parametric bootstrap to quantify the variability of the estimated causal structures. We applied these methods to investigate how the high mobility group AT-Hook 2 (HMGA2) gene is incorporated in the protein 53 signalling pathway playing an important role in head-and-neck squamous cell carcinoma. Our results were quite stable and found direct associations between HMGA2 and other relevant proteins, but they did not provide clear support for the claim that HMGA2 itself is a key regulator gene.

Address for correpondence: Ronja Foraita, Leibniz Institute for Prevention Research and Epidemiology—BIPS, Achterstrasse 30, Bremen 28359, Germany.
E-mail: foraita@leibniz-bips.de

1. Introduction

The analysis of gene regulatory networks aims to improve our understanding of the relationships between genes in biological processes that are associated with disease (van Dam *et al.*, 2018). The notion of a 'regulatory network' is rather vague, and a wide variety of methods, based on different principles, have been suggested and applied (Husmeier *et al.*, 2006; Albieri and Didelez, 2014).

Typically, gene regulatory networks are meant to represent biological interactions between simultaneously or non-simultaneously expressed genes but as such they do not necessarily reflect or allow statements about causal relations (Bansal *et al.*, 2007). Regulator genes, especially those which are involved in cell growth or cell cycle regulation, are potential targets for drug development, such as EGFR in lung cancer patients (Liu *et al.*, 2017). With a view to such potential interventions it is crucial to use methods that are specifically designed for analysing the causal structure underlying gene regulation, as opposed to merely investigating their associations.

At the intersection of statistics and artificial intelligence, we find various methods and algorithms, known as causal discovery, which aim to identify causal relationships from observational or (partially) experimental data (Spirtes *et al.*, 2000; Zhang *et al.*, 2018; Heinze-Deml *et al.*, 2018; Spirtes and Zhang, 2018). These define causal relationships explicitly in terms of intervention effects, e.g. gene knock-outs: A would be causal for B if intervening in A affects the distribution of B. Maathuis *et al.* (2010) demonstrated with an application to yeast gene expressions that causal discovery algorithms outperformed traditional statistical methods based on regression and prediction such as the elastic net. Causal discovery essentially assumes that the underlying causal structure can be represented by a directed acyclic graph (DAG) on measured, and possibly additional latent, variables, e.g. gene expressions. (Note that we use the terms network and graph interchangeably.) The aim then is to identify the most plausible DAG, given the data, under specific assumptions allowing a causal interpretation. Numerous algorithms have been proposed for this; these are mainly versions of either so-called constraint-based or score-based approaches. Constraint-based methods match conditional independences that are found in the data with those implied by a DAG, whereas score-based methods select the most plausible DAG on the basis of a score assessing the fit of a DAG to the data. For examples of the use of causal discovery in the context of gene regulation see Chu *et al.* (2003), Sachs *et al.* (2005), Opgen-Rhein and Strimmer (2007) and Maathuis *et al.* (2009).

Here, we are interested in applying causal discovery to investigate the causal relationships between high mobility group AT-Hook 2 (HMGA2), a gene that in adult tissues is expressed only in both malignant and benign tumour formation, and some target genes of the p53 signalling pathway, an important cell signalling pathway involved in the carcinogenesis of many cancers such as head-and-neck squamous cell carcinoma (HNSCC). It has been suggested that, within the p53 signalling pathway, the HMGA2 gene is a key regulator gene (Wei *et al.*, 2010). However, empirical evidence to support this hypothesis has been obtained only from cell culture studies, but not in HNSCC patients (Ji *et al.*, 2008). We address this question by analysing gene expression data of tumour tissues from $n = 208$ HNSCC patients.

In our investigation, we needed to address certain methodological challenges. First, our data were incomplete in that individual measurements were missing for 24% of the patients. Analysis of only the complete cases (CCs) would ignore much information and is well known potentially to induce bias (Sterne *et al.*, 2009). Standard methods for dealing with missing data are multiple

imputation (van Buuren, 2018) or the expectation–maximization (EM) algorithm (Dempster *et al.*, 1977). Although estimation in graphical models (Didelez and Pigeot, 1998) and score-based search algorithms have been combined with EM before (Friedman, 1997; Scutari, 2010), we here propose a first idea of how multiple imputation can be combined with constraint-based causal discovery. We consider this a particularly promising avenue as multiple imputation methods are easy to use because there is plenty of software which can handle a wide spectrum of data situations; at the same time constraint-based causal discovery algorithms are very flexible and are used increasingly in biomedical fields of application. Second, causal discovery in general relies on strong assumptions, e.g. some algorithms assume the absence of unobserved confounding; moreover, it can be unstable and sensitive for instance regarding the settings of tuning parameters. Hence, we carefully carried out various sensitivity analyses, e.g. allowing for latent confounding, and assessed the robustness of our results by bootstrapping the selected causal graphs (Friedman *et al.*, 1999; Pigeot *et al.*, 2015).

1.1. *Background on the protein 53 signalling pathway*

The tumour suppressor gene TP53 is an important anticancer gene because of its frequent mutations in most human solid cancers such as HNSCC (Stewart and Wild, 2014; Parameswaran and Burtness, 2018). The gene TP53 encodes the p53 tumour suppressor protein which initiates cell cycle arrest and apoptosis in response to cellular deoxyribonucleic acid (DNA) damage. Inactivation of the p53 signalling pathway caused by genetic alterations of TP53 is the most frequent event in HNSCC and has been attributed to tobacco smoking and alcohol consumption (Stewart and Wild, 2014; Peltonen *et al.*, 2010). Wei *et al.* (2010) suggested that the oncofetal stem cell factor HMGA2 is a key regulator gene of the tumour protein 53 (TP53). HMGA2 is expressed during early embryogenesis and in cell differentiation. In adult tissues it is expressed in both malignant and benign tumours of different sites including HNSCC (Miyazawa *et al.*, 2004; Klemke *et al.*, 2009; Hetland *et al.*, 2012; Piscuoglio *et al.*, 2012). However, the exact role of HMGA2 in tumorigenesis remains unclear. In cell culture studies, HMGA2 has been reported to be associated with the target genes CDKN2A/p14, MDM2, CDKN1A/p21 and BAX of the p53 signalling pathway (Markowski *et al.*, 2010, 2011). The encoded proteins of these four genes are key players in the p53 pathway. Hence, our analysis focuses on these five genes, although the p53 pathway itself is composed of hundreds of genes, many of which may also be relevant to HNSCC (Levine *et al.*, 2006).

The tumour suppressor protein p21, encoded by the senescence gene CDKN1A, is a major target of p53 activity that induces cell cycle arrest to repair DNA damages (Vogelstein *et al.*, 2000). If the damage is too serious, BAX is expressed in response to the tumour suppressor p53 to induce apoptotic cell death (Hanahan and Weinberg, 2011). Gavathiotis *et al.* (2012) discovered that pharmacologic activation of apoptosis in cancer cells by triggering BAX is possible. MDM2 is an important negative regulator protein of p53 that enhances cancer growth. MDM2 and p53 are connected through an autoregulatory feedback loop that maintains low cellular p53 levels in the absence of stress (Moll and Petrenko, 2003). Its amplification frequency in HNSCC is overexpressed in up to 46% of cases (Millon *et al.*, 2001) leading to inhibition of p53 (Parameswaran and Burtness, 2018). CDKN2A produces the protein p14ARF: a tumour suppressor that reduces MDM2 and stabilizes p53 (Parameswaran and Burtness, 2018).

Although the main aim of our study is to investigate the role of HMGA2 based on a sample of head-and-neck cancer tumour tissue specimens, the specific question of human-papilloma-virus (HPV) related positive HNSCC is of great interest. HPV positive subtypes show different molecular patterns such as higher expressions of the CDKN2A encoded tumour suppressor protein p14/p16 and lower mutation rates of TP53 (Faraji *et al.*, 2018). We therefore conduct a

secondary analysis on an HPV positive subsample to investigate whether HMGA2 might have a different role in HPV-induced HNSCC.

2. Methods

Our investigation of the causal structure underlying the p53 pathway required the combination of causal discovery methods with missing data methods, followed by specific sensitivity analyses. Some background on these methods, that is necessary to understand our results, is presented in this section.

2.1. Causal graphs

We give a brief introduction to (causal) graphs and the most relevant search algorithms. In particular we focus on constraint-based algorithms but, for comparison, we also consider a score-based algorithm.

2.1.1. Graph terminology

A graph $\mathcal{G} = (\mathbf{V}, \mathbf{E})$ combines a set of vertices (nodes) $\mathbf{V} = \{X_1, \ldots, X_p\}$ and a set of edges $\mathbf{E} \subseteq \mathbf{V} \times \mathbf{V}$. The latter can be of different types: directed ($\rightarrow$, $\leftarrow$), undirected (–) or bidirected ($\leftrightarrow$) edges. Two vertices are said to be *adjacent*, if they are connected by an edge. If $X_i \rightarrow X_j$ is in $\mathcal{G}$, then X_i is a *parent* of X_j and X_j is a *child* of X_i. A *path* is a sequence of distinct adjacent vertices. On a *directed path* all edges point in the same direction. A directed *cycle* is a directed path that starts and ends at the same vertex. A v-structure in $\mathcal{G}$ is an ordered triple (X_i, X_j, X_k) such that $\mathcal{G}$ contains the edges $X_i \rightarrow X_j$ and $X_j \leftarrow X_k$ and X_i and X_k are not adjacent in $\mathcal{G}$.

A graph that contains only *directed* edges is called a *directed* graph and a graph that contains *directed* and *undirected* edges is called a *partially directed* graph. The *skeleton* of a (partially) directed graph is the undirected version of this graph after all arrowheads have been removed. A (*P*)*DAG* is a (partially) directed graph without directed cycles.

2.1.2. Probabilistic and causal interpretation

Graphs are used to encode the conditional independence structure of a multivariate distribution (Lauritzen, 1996), and when interpreted causally they additionally represent the effects of interventions on one or more nodes (Pearl, 2009; Didelez, 2018).

Let the vertices of a DAG correspond to random variables $X_1, \ldots, X_p$. Then we can read off conditional independences via the graphical criterion known as *d*-separation (Pearl, 2009), and, equivalently, a joint probability distribution P factorizes according to the DAG $\mathcal{G}$, if this joint distribution over $\mathbf{V} = \{X_1, \ldots, X_p\}$ can be written as $P(\mathbf{V}) = \Pi_{i=1}^{p} P\{X_i | \mathrm{pa}(X_i)\}$, where $\mathrm{pa}(X_i)$ denotes the set of parents of X_i. Moreover, the probability distribution P is said to be *faithful* to the DAG $\mathcal{G}$ if *every* conditional independence in P is implied by the *d*-separations in $\mathcal{G}$. Under faithfulness, two adjacent variables in a DAG remain associated even after conditioning on any subset of the other variables in the graph. Hence, in what follows we shall say that two adjacent variables have a direct association; but note that 'direct' is relative to the particular set of variables considered. Non-adjacent variables connected by an open path are marginally, i.e. indirectly, associated. Interpreted causally, a DAG implies that an intervention setting an arbitrary variable X_i to $\tilde{x}_i$ corresponds to replacing the factor $P\{X_i | \mathrm{pa}(X_i)\}$ in the above factorization by the indicator function $I(X_i = \tilde{x}_i)$ whereas all other factors remain the same. In particular, this can be plausible only if the vertices $\mathbf{V}$ include all common causes of any two variables in the graph. In other words, we would need to assume that there is no latent (i.e. unobserved) confounding of any pair of observed variables. This assumption is known as *causal sufficiency*.

The estimation of a DAG from observational data is hampered by the fact that different DAGs can be *Markov equivalent*, i.e. encode the same *d*-separations and hence the same conditional independences. This means that certain causal structures cannot be distinguished on the basis of their implied independence structure, and hence we can identify only the set of equivalent graphs. For instance, $X_i \rightarrow X_j \rightarrow X_k$ and $X_i \leftarrow X_j \leftarrow X_k$ both imply that X_i and X_k are independent given X_j, so without further information we cannot say whether X_i is a cause or an effect of X_j and X_k. Such further information could be provided by including more variables and/or subject matter background knowledge. Markov equivalent DAGs share the same skeleton and the same v-structures. The corresponding Markov equivalence class can uniquely be represented by a *completed partially directed acyclic graph* (CPDAG) (Chickering, 2002). A CPDAG is a partially directed graph where a directed edge means that this directed edge is present in all DAGs in the Markov equivalence class, whereas an undirected edge X_i–X_j means that there is at least one DAG in the equivalence class with $X_i \rightarrow X_j$, and at least one DAG with $X_i \leftarrow X_j$. For the above example, the CPDAG would be X_i–X_j–X_k, which additionally contains $X_i \leftarrow X_j \rightarrow X_k$, but not the v-structure $X_i \rightarrow X_j \leftarrow X_k$. The latter implies only the marginal independence of X_i and X_k and has no other equivalent DAG on (X_i, X_j, X_k).

A more general type of graphs than causal DAGs, relaxing causal sufficiency, are *maximal ancestral graphs* (MAGs) (Richardson and Spirtes, 2002). In full generality, MAGs also allow for selection on latent variables but we shall not make use of this aspect here. MAGs contain directed and bidirected edges, but no directed cycles. Under the assumption of faithfulness, an MAG encodes all and only those conditional independence relationships over the observed variables that are also satisfied by an underlying causal DAG over the same observed and further unobserved (latent) variables. The conditional independence relationships can be read off an MAG via the *m*-separation criterion, which is a generalization of *d*-separation. Similarly to DAGs, different MAGs can be Markov equivalent. The corresponding Markov equivalence class can be represented by a *partial ancestral graph* (PAG) (Zhang, 2008a). Thus, a PAG preserves the causal features of all DAGs that share the same set of observable conditional independence statements and ancestral relationships, without making any restrictions on the number of unobserved confounding variables.

PAGs contain different types of edges ($\rightarrow$, $\leftrightarrow$, $\circ\!\!-\!\!\circ$ and $\circ\!\!\rightarrow$) and edge marks: arrowhead '>', tail '–' and circle '∘'. An arrowhead or tail means respectively that this arrowhead or tail is present in all MAGs of the equivalence class, whereas a circle means that there are at least two MAGs in the equivalence class where the edge mark is at least once an arrowhead and otherwise a tail. Bidirected edges $X_i \leftrightarrow X_j$ represent latent confounding and mean that neither X_i is a cause of X_j nor X_j is a cause of X_i; see Zhang (2008a) for more information on the interpretation of MAGs and PAGs.

All of these different graphs share two interpretations:

(a) every missing edge corresponds to a conditional independence relationship, and
(b) only the edge $X_i \rightarrow X_j$ uniquely implies that X_i is a cause of X_j and that X_j is not a cause of X_i.

2.2. Causal discovery from observational data

Numerous causal discovery algorithms exist aiming at estimating the causal structure of the underlying data-generating mechanism (Maathuis and Nandy, 2016; Kalisch and Bühlmann, 2014; Spirtes and Zhang, 2018). In this paper, we focus on two constraint-based and one score-based algorithms which we briefly introduce here.

2.2.1.   *Constraint-based algorithms*

Constraint-based methods for causal discovery first establish conditional independences from the data and then construct a causal structure that agrees with these conditional independence constraints. The main idea behind this approach is that if two variables X_i and X_j are not adjacent in the underlying causal DAG, i.e. neither X_i is a direct cause of X_j nor vice versa, then they must be conditionally independent given *some* subset of the remaining variables. Causal discovery turns this around and looks for conditional independences to infer separations, relying on the faithfulness assumption. As mentioned above, conditional independence information, alone, only allows us to infer Markov equivalence classes of causal graphs (CPDAGs or PAGs) which may still be useful and allow novel insights into the underlying causal structure. More specifically, constraint-based methods employ a series of conditional independence tests as their starting point. Although in principle this could be an infeasibly large number of tests, one for each pair of variables and every possible separating set, in practice we find that either if the number of variables is not too large, or if the graph is sparse (Kalisch and Bühlmann, 2007), then the algorithms still terminate in acceptable time. This is because constraint-based algorithms start by considering small separating sets and for moderately sparse graphs it is likely that they can terminate before considering larger separating sets. When, as in our case, all variables are continuous, it is common to test for zero partial correlations. Such a test implies conditional independence under a joint multivariate Gaussian distribution of the random variables but is informative in its own right as testing for absence of linear dependences (Cox and Wermuth, 1996).

In Section 2.3, we extend a partial correlation test to account for multiply imputed data and show how this can be incorporated into constraint-based causal discovery algorithms. In particular, we combined PC stable and fast causal inference (FCI) stable search with multiple imputation and applied these to our data.

2.2.2.   *PC stable algorithm*

The PC algorithm (Spirtes *et al.*, 2000) is the most prominent algorithm for causal discovery from observational data. Under the assumptions that the probability distribution is faithful to a DAG and under causal sufficiency (no latent confounding) it recovers the true CPDAG when provided with the correct conditional independence constraints. In practice, however, statistical tests make mistakes which lead to various problems for the PC algorithm. Under certain distributional assumptions, e.g. multivariate Gaussianity, it can be shown that the PC algorithm still consistently selects the true CPDAG (Kalisch and Bühlmann, 2007; Maathuis and Nandy, 2016), whereas uniform consistency is problematic (Robins *et al.*, 2003). For finite samples, the result of the original algorithm can depend on the order in which the variables are entered, but this can be modified to be order independent (Colombo and Maathuis, 2014). Here, we shall use the fully order-independent version and call it the PC stable algorithm. The PC algorithm follows three main steps that are reviewed in more detail in Kalisch and Bühlmann (2014).

Step 1: the skeleton of the DAG is estimated by performing a series of conditional independence tests for each pair of variables.
Step 2: some edge directions can then be determined by identifying v-structures from the skeleton and the conditional independences.
Step 3: further edge directions can be determined logically based on the partially directed graph from step 2, as no additional v-structures or cycles are allowed.

2.2.3.   *Fast causal inference stable algorithm*

Relaxing the assumption of causal sufficiency, the FCI algorithm (Spirtes *et al.*, 2000) allows

causal discovery with latent variables. Although this is often more realistic, it comes at the price of a typically much more difficult to interpret and vague result. Given correct conditional independence information, the FCI algorithm recovers the true PAG under the assumption that the observed probability distribution is faithful to a DAG containing the observed as well as unobserved variables. Consistency can be shown in high dimensional settings (Colombo *et al.*, 2012). In practice, problems similar to those for the PC algorithm occur. The FCI stable algorithm is again modified so as to be fully order independent (Colombo and Maathuis, 2014). The FCI algorithm proceeds as follows (see for example Kalisch and Bühlmann (2014) for more details).

Steps 1 and 2 are analogous to those of the PC algorithm.
Step 3: update the skeleton after computing so-called *possible D-SEP* sets (Spirtes *et al.*, 2000; Colombo *et al.*, 2012) and test edges in the initial skeleton for conditional independence given subsets of possible D-SEP sets. This might lead to edge removals in the skeleton and extensions of separating sets.
Step 4: renew determination of v-structures.
Step 5: apply Zhang's 10 orientation rules (Zhang, 2008b).

2.2.4. *Score-based algorithms*

Score-based causal discovery algorithms assign a score to each candidate DAG and aim at finding the DAG with the optimal score. The score function is usually chosen to be score equivalent, so that the same score is assigned to DAGs of the same Markov equivalence class. Typically, the score is likelihood based with some penalty for complexity of the graph, for instance the Bayesian information criterion (BIC) (Chickering, 1995). Hence, score-based algorithms require the full specification of a likelihood, whereas constraint-based algorithms rely only on suitable statistical tests for conditional independence. Most score-based approaches assume causal sufficiency as it is computationally expensive to search among models that allow for latent confounding. Generally, the main difficulty for score-based algorithms is to ensure that the search space is visited in a manner ensuring that, ideally, the global optimum is found. Greedy equivalence search ensures that a global optimum is found (Chickering, 2002), but other greedy search strategies can perform better in practice (Gillispie and Perlman, 2002).

In case of incomplete data, score-based algorithms lend themselves to be combined with the EM algorithm as they both rely on the likelihood. However, under general incomplete-data patterns the likelihood does not factorize in the same way as for complete data so brute force combination of EM and greedy search would be computationally very expensive as it cannot be carried out locally. In our application, we compared our multiple-imputation approach for constraint-based causal discovery with a suggestion of Friedman (1997) for combining greedy search algorithms with the EM algorithm as implemented in `bnlearn` (Scutari, 2010) which we call the structural EM (SEM) algorithm. The key idea, here, is that the most plausible graph is determined within each iteration of the EM algorithm, i.e. only for the current values of the parameters. The practical efficiency of the SEM algorithm has been demonstrated empirically in a variety of settings (Friedman, 1997).

2.3. *Multiple imputation for constraint-based causal discovery*

Multiple imputation is a widely used flexible technique for handling missing values. It creates $M > 1$ complete data sets where missing values are filled in by plausible, typically model-based, values. The imputation models can be motivated by a Bayesian approach or by fully conditional specification of the joint distribution (van Buuren, 2018). Each of these M data sets is analysed

separately by the standard complete-data statistical procedure and then the *M* results are pooled into an overall estimate, and appropriate standard errors are calculated according to Rubin's rules (Rubin, 2004). Multiple imputation relies on the data being missing at random, i.e. which observations are missing is independent of the actual missing values given the observed data (Little and Rubin, 2002). The plausibility of the missingness at random (MAR) assumption can be improved by including suitable predictors in the imputation model, especially also predictors that are not used in the analysis model. We discuss this in the context of our application in Section 3.1.

Multiple imputation and Rubin's rules, as described above, are designed for parameter estimation. Our aim, here, is causal discovery, and the basis of constraint-based algorithms is not parameter estimates but a series of conditional independence tests. Instead of a reliable pooled estimated parameter value we wish to construct a reliable pooled test decision for each conditional independence test required by the chosen causal discovery algorithm.

The key procedure that we propose is as follows.

Step 1: apply multiple imputation as usual and appropriate to create *M* data sets.
Step 2: on each of the *M* data sets compute the desired test statistics for testing independences of variables X_i and X_j given a variable set S_l as required by the chosen constraint-based algorithm.
Step 3: combine the *M* test statistics by using Rubin's rules and input the resulting pooled test decision into the next steps of the chosen constraint-based algorithm.

Below, we illustrate this procedure with the special case of a test for zero partial correlation. This is the standard conditional independence test that is used by constraint-based algorithms when a multivariate Gaussian distribution seems appropriate for the data at hand. The procedure was then combined with the PC stable and FCI stable algorithms when applied to our data on the p53 pathway.

2.3.1. *Partial correlation and Fisher's z-transformation*

Let $\{X_1, \ldots, X_p\} \in \mathbb{R}^p$ be random variables corresponding to the vertices in **V**. The partial correlation coefficient between two variables X_i and X_j given the subvector $\mathbf{X}_{S_l} = (X_{k_1}, \ldots, X_{k_l})$, $S_l = \{k_1, \ldots, k_l\} \subset \mathbf{V} \setminus \{i, j\}$, is the correlation between X_i and X_j that remains after adjusting for the linear effects of $\mathbf{X}_{S_l}$. Letting $\mathbf{Z} = (X_i, X_j)^{\mathsf{T}}$, the partial covariance matrix is derived from the partitioned covariance matrix

$$\operatorname{cov}\begin{pmatrix} \mathbf{Z} \\ \mathbf{X}_{S_l} \end{pmatrix} = \begin{pmatrix} \Sigma_{\mathbf{Z},\mathbf{Z}} & \Sigma_{\mathbf{Z},\mathbf{X}_{S_l}} \\ \Sigma_{\mathbf{X}_{S_l},\mathbf{Z}} & \Sigma_{\mathbf{X}_{S_l},\mathbf{X}_{S_l}} \end{pmatrix}$$

by using the decomposition

$$\Sigma_{\mathbf{Z}|\mathbf{X}_{S_l}} = \Sigma_{\mathbf{Z},\mathbf{Z}} - \Sigma_{\mathbf{Z},\mathbf{X}_{S_l}} \Sigma_{\mathbf{X}_{S_l},\mathbf{X}_{S_l}}^{-1} \Sigma_{\mathbf{X}_{S_l},\mathbf{Z}}.$$

The pairwise partial correlation of X_i and X_j given $\mathbf{X}_{S_l}$ obtains as

$$\rho_{ij|S_l} = \frac{\sigma_{ij|S_l}}{\sqrt{(\sigma_{ii|S_l} \sigma_{jj|S_l})}},$$

where $\sigma_{ij|S_l}$ are elements of the 2×2 partial covariance matrix $\Sigma_{\mathbf{Z}|\mathbf{X}_{S_l}} = \{\sigma_{ij|S_l}\}$.

Under the assumption that the random vector $(X_i, X_j, X_{k_1}, \ldots, X_{k_l})^{\mathsf{T}}$ follows a multivariate Gaussian distribution, $\rho_{ij|S_l} = 0$ if and only if X_i and X_j are conditionally independent given $\mathbf{X}_{S_l}$. This is used to determine the presence or absence of a potential edge in $\mathcal{G}$. Although the empirical

partial correlation coefficient $\hat{\rho}_{ij|S_l}$ would be an obvious test statistic, it has the drawback that it is not normally distributed under the null hypothesis (Hotelling, 1953). Fisher (1924) suggested transforming the partial correlation coefficient into the z-statistic

$$z_{ij|S_l} = \frac{1}{2} \ln \left(\frac{1 + \hat{\rho}_{ij|S_l}}{1 - \hat{\rho}_{ij|S_l}} \right), \tag{1}$$

which is approximately normally distributed with variance $1/(n-l-3)$, where n is the sample size and $l = |S_l|$. Under the null its asymptotic distribution has mean 0.

2.3.2. Fisher's z-test with multiple imputation

To test for zero partial correlation by using multiple imputation we follow the suggestion by D'Angelo *et al.* (2012). For each of the M data sets obtained by multiple imputation, we compute Fisher's z-transformations $z_{ij|S_l}^{(m)}$, $m = 1, \ldots, M$, and its variance with equation (1). These M coefficients and their variances are then combined into one multiple-imputation inference by using Rubin's rules as follows.

Let $z_{ij|S_l}^{(m)}$ be the z-transformation of the partial correlation for the mth imputed data set testing the conditional independence between variables X_i and X_j given a variable subset $\mathbf{X}_{S_l}$. The pooled test statistic is the average of the M individual statistics

$$\bar{z}_{ij|S_l} = \frac{1}{M} \sum_{m=1}^{M} z_{ij|S_l}^{(m)}.$$

The total variance estimator

$$V_{ij|S_l} = \bar{W}_{ij|S_l} + \left(1 + \frac{1}{M} \right) B_{ij|S_l}$$

is a weighted sum of the within variance $\bar{W}_{ij|S_l}$, i.e. the average of the complete-data sample variances, and the variance $B_{ij|S_l}$ between the M completed data sets:

$$\bar{W}_{ij|S_l} = \frac{1}{M} \sum_{m=1}^{M} \hat{W}_{ij|S_l}^{(m)} = \frac{1}{M} \sum_{m=1}^{M} \frac{1}{n-l-3} = \frac{1}{n-l-3}$$

$$B_{ij|S_l} = \frac{1}{M-1} \sum_{m=1}^{M} (z_{ij|S_l}^{(m)} - \bar{z}_{ij|S_l})^2.$$

Finally, the conditional independence desired can now be tested with multiple-imputed data, using the combined test statistic

$$T = \frac{\bar{z}_{ij|S_l}}{\sqrt{V_{ij|S_l}}} \tag{2}$$

which, under the null, has a Student t-distribution with ν degrees of freedom. Commonly, a Satterthwaite approximation (Little and Rubin (2002), page 87) is used to calculate ν as

$$\nu = (M-1) \left(1 + \frac{M}{M+1} \frac{\bar{W}_{ij|S_l}}{B_{ij|S_l}} \right)^2.$$

To implement the above approach in practice, we modified the relevant functions of the PC and FCI stable algorithm included in the R package `pcalg` (Kalisch *et al.*, 2012) to be used for multiple-imputed data. In what follows, we shall refer to them as PC-MI and FCI-MI. Our R functions can be downloaded from `https://github.com/bips-hb/micd`.

2.4.   Sensitivity analyses

In our investigation of the p53 pathway, we regard the PC-MI algorithm as a first step. We considered the following additional analyses to assess the sensitivity towards the different assumptions.

2.4.1.   Relaxing causal sufficiency

The PC algorithm relies crucially on the assumption of causal sufficiency. This means that all common causes of two or more measured variables must also be measured and taken into account. This is in many situations an unrealistic assumption; for instance in our application we were aware that many more proteins and biomarkers are known to be relevant to the p53 pathway than those five that were available for our analysis. Hence we use the FCI stable algorithm, combined with the same multiple-imputation process, as an alternative. Note that, in the presence of latent variables, the MAR assumption is still sufficient for multiple imputation to rely on measured variables only.

2.4.2.   Score-based search and expectation–maximization algorithm

Both PC and FCI algorithms are based on conditional independence tests. This approach can be criticized: a statistical test can be wrong in both directions, i.e. include an edge where there should be none or vice versa. Especially in situations of low power, erroneous decisions are particularly detrimental for constraint-based search which relies on finding *independences*. If errors are made early on during the algorithm this could have very adverse effects on the result. A basic sensitivity analysis should therefore vary the nominal significance level that is used for the test decision. Moreover, it is prudent to consider an alternative method that takes a very different approach, such as a score-based algorithm. Relying on a likelihood-based score, such an algorithm essentially evaluates the plausibility of the whole given DAG jointly instead of one edge at a time; for example it can avoid conflicting edge orientations. Moreover, the score-based approach enabled us to compare multiple imputation with a different way of handling incomplete data: the SEM algorithm (Friedman, 1997).

2.4.3.   Testwise deletion and complete-case analysis

As multiple imputation is a computationally demanding method and relies on the correctness of the imputation model, we compare our results with two simpler methods for dealing with missing values: CC analysis (which is also known as listwise deletion) and testwise deletion (TD) (which is also known as pairwise deletion or available case analysis). In CC analysis a case including a missing value is entirely deleted from the data. TD omits only cases with missing values in those variables required for the current conditional independence test when performing the PC or FCI algorithm. Both methods are consistent under missingness completely at random but not under MAR (van Buuren (2018), page 9). Very recently, Tu *et al.* (2019) have suggested a correction to TD for constraint-based search so that it is valid under MAR extending the results of Strobl *et al.* (2018), but to our knowledge this is not implemented in any R package, yet. Both approaches are less efficient than multiple imputation as they do not use all the data, though TD improves on CC analysis. However, in contrast with multiple imputation they do not require any imputation models and they do not make use of external data. Strobl *et al.* (2018) showed that TD for FCI outperforms CC analysis even under certain missingness not at random mechanisms.

2.4.4.   Assessing the uncertainty of selected graphs

Although it is common to indicate the uncertainty or variability of an estimate by computing

standard errors or confidence intervals, it is not straightforward to quantify the uncertainty when the estimated object is a graph. In our analysis, we followed the suggestions of Friedman *et al.* (1999) and Pigeot *et al.* (2015) and used a non-parametric bootstrap analysis of the whole selection process including multiple imputation within the bootstrap. This means that for each of a number of bootstrap samples a graph or an equivalence class of graphs is estimated; if these graphs are very similar we can say that the selection process is stable. Hence we need to assess the similarity of graphs for which we use the measures based on the (structural) Hamming distance (Hamming, 1950; Tsamardinos *et al.*, 2006) and performance measures such as precision and recall.

3. Application to the p53 signalling pathway

Our study aimed to investigate the causal relationships between HMGA2 and four specific genes, CDKN2A/p14, MDM2, CDKN1A/p21 and BAX, as part of the p53 pathway, by applying causal discovery methods. As HMGA2 is exclusively expressed in tumour and embryonic tissues, we analysed only HNSCC cases that were recruited in a multicentre study: 'Alcohol related cancers and genetic susceptibility in Europe' (Lagiou *et al.*, 2009).

3.1. The data

Tumour tissues (formalin fixed and paraffin embedded) from 208 patients with histologically confirmed HNSCC were collected in Bremen between 2003 and 2005 (Friemel *et al.*, 2016). Tumour sites of the upper aerodigestive tract comprised the oral cavity, tonsils, pharynx and larynx. Cases with *in situ* carcinoma and oesophageal cancer were not included. Information about risk factors and covariates, including tobacco smoking and alcohol drinking, was assessed through standardized computer-assisted personal interviews and diagnoses were confirmed by histology. Tumour stage was classified according to the Union for International Cancer Control (tumour–node–metastasis (TNM) stage I–IV). If one of the mandatory grading parameters T (primary tumour), N (regional lymph nodes) and M (distant metastasis) was not reported, tumour stage was unknown.

3.1.1. Laboratory analyses

Ribonucleic acid (RNA) isolation from formalin-fixed and paraffin-embedded tissue samples and complementary DNA synthesis of all five genes followed by relative quantification of transcription levels by realtime polymerase chain reaction (PCR) assays were performed in triplicate with a special primer for CDKN2A/p14 as described by Markowski *et al.* (2011). The house keeping gene HPRT served as endogenous normalizer for quantification of gene expression levels (Markowski *et al.*, 2010; Lallemant *et al.*, 2009). Gene expression values were base 2 logarithmically transformed and standardized with mean 0 and variance 1. Expression levels that were more distant than three interquartile ranges from the upper or lower quartile were omitted as outliers.

In a subset of cases ($n = 187$) HPV DNA was detected by using the primer system GP5+/6+ developed by de Roda Husman *et al.* (1995) To prevent contamination the PCR Core Kit-PLUS (Roche) was adapted to the PCR. HPV type-specific PCR and primers were used as control to verify the results for HPV-16 (ATATAAGGGGTCGGTGGACCG and GCAATG-TAGGTGTATCTCCATGC) and HPV-18 (AAGGATGCTGCACCGGCTGAA and CACG-CACACGCTTGGCAGGTTT). P16 immunohistochemistry, expression analysis and scoring were performed as described by D'Souza *et al.* (2016).

Table 1. Clinical and pathological characteristics of HNSCC grouped by availability of genetic information†

Characteristic		CGI (n=159)		IGI (n=49)		Total (n=208)	
		n	%	n	%	n	%
Sex	Female	24	15.1	10	20.4	34	16.3
	Male	135	84.9	39	79.6	174	83.7
Age	Mean (±SD)	58.2	8.6	58.5	9.2	58.3	8.7
Tumour site	Oral cavity (C01–C06)	55	34.6	9	18.4	64	30.8
	Tonsils (C09)	21	13.2	6	12.2	27	13.0
	Pharynx (C10–C13)	38	23.9	19	38.8	57	27.4
	Larynx (C32)	45	28.3	15	30.6	60	28.8
Union for International	I	14	8.8	1	2.0	15	7.2
Cancer Control stage	II	22	13.8	4	8.2	26	12.5
based on TNM	III	15	9.4	6	12.2	21	10.1
	IV	71	44.7	25	51.0	96	46.2
	X (missing)	37	23.3	13	26.5	50	24.0
TNM classification	T missing	13	8.2	2	4.1	15	7.2
	N missing	29	18.2	6	12.2	35	16.8
	M missing	61	38.4	14	28.6	75	36.1
Therapy	Surgery	130	81.8	42	85.7	172	82.7
	Chemotherapy	65	40.9	19	38.8	84	40.4
	Radiotherapy	107	67.3	37	75.5	144	69.2
Hospital of tumour resection	North	38	23.9	17	34.7	55	26.4
	Centre 1	31	19.5	8	16.3	39	18.8
	Centre 2	90	56.6	24	49.0	114	54.8
HPV type 16/18	Negative	123	77.4	40	81.6	163	78.4
	Positive	21	13.2	3	6.1	24	11.5
	Missing	15	9.4	6	12.2	21	10.1
P16 status	Negative	106	66.7	26	53.1	132	63.5
	Positive	35	22.0	10	20.4	45	21.6
	Missing	18	11.3	13	26.5	31	14.9
Smoking status	Current	120	75.5	42	85.7	162	77.9
	Never or former	39	24.5	7	14.3	46	22.1
Pack-years	Median (MAD)	36.5	20.6	42	22.2	37.2	21.1
	Missing	3	1.9	0	0	3	1.4
Alcoholic drinks per day	Median (MAD)	1	1.2	1.6	2	1.2	1.3
	Missing	3	1.9	0	0	3	1.4
Education	< 10 years	106	66.7	39	79.6	145	69.7
	⩾ 10 years	53	33.3	10	20.4	63	30.3
Missing gene expression data							
CDKN2A/p14	Missing	—	—	23	46.9	23	11.1
BAX	Missing	—	—	5	10.2	5	2.4
HMGA2	Missing	—	—	7	14.3	7	3.4
MDM2	Missing	—	—	25	51.0	25	12.0
CDKN1A/p21	Missing	—	—	7	14.3	7	3.4

†Unless otherwise stated, values are frequencies *n* and percentages. CGI, complete gene information; IGI, incomplete gene information; TNM classification of malignant tumours: T describes the primary tumour, N the lymph nodes involved and M the distant metastasis.

3.1.2. Patients characteristics

208 HNSCC cases were included in the analysis (Table 1). The male–female ratio was 5:1, the mean age was 58 years (±8.7 standard deviations (SDs)). 127 patients (61%) had localized tumours (T1 or T2). Lymph node involvement (N1 or N2) was reported in 52% of the cases. A subset of 24 of 187 patients with available test results for high-risk HPV 16/18 were positive. P16 protein expression was detected in tumour tissues of 22% and was associated with PCR test

results for high risk HPV 16/18 ($\chi^2 = 53.0$; $p < 0.001$). Subjects reported a median number of pack-years of 37 (median absolute deviation MAD ± 21) and 162 (78%) patients were classified as current smokers. 114 (55%) patients reported that they consumed at least one alcoholic drink per day.

Overall, 49 HNSCC cases (24%) had missing gene expression values for at least one gene. Gene expression of CDKN2A/p14 and MDM2 were most frequently missing in 11% and 12% of all patients respectively. No particular missingness pattern was observed between gene expression values and other variables.

3.1.3. *Missing values*
Missing gene expression values are generally caused by fragmented RNA in processed samples where longer amplification products are more likely to be affected. Fragmentation occurs spontaneously in a way that possibly depends on how the sample was processed after surgical tumour resection, but it is thought not to be related to the expression itself. This suggests that the missingness is likely to be completely at random; but to be on the safe side we assume MAR given the following variables that are always observed: hospital, age, sex, tumour site, received therapies (surgery, chemotherapy or radiotherapy), smoking status, education time, survival time (in months) and censoring status, as well as the observed values of the remaining variables: the other gene expressions, TNM classification, HPV status, p16 status, pack-years of tobacco smoking and alcoholic drink consumption per day.

3.2. *Specification of statistical methods*
3.2.1. *Multiple imputation*
With the justification given above, we assume that the unobserved data are missing at random given the named variables so that these are used as predictors in the imputation models. As imputation approach we applied a fully conditional specification for multiple imputation via chained equations as implemented in the algorithm MICE (van Buuren and Groothuis-Oudshoorn, 2011). MICE imputes data variable by variable by specifying a scale-specific imputation model for each variable. Missing values were imputed $M = 100$ times.

3.2.2. *Causal discovery algorithms*
We applied PC-MI so that it produced a fully order-independent output and FCI-MI so that conflicting edge orientations were resolved by majority rule (Colombo and Maathuis, 2014) as implemented in the R package `pcalg` (Kalisch *et al.*, 2012). For both constraint-based algorithms we used a nominal significance level of 5% for each conditional independence test. This value is typically treated as a tuning parameter that regulates the edge density, where smaller α-values lead to sparser graphs. In our sensitivity analysis we also used the alternative α-values 1% and 10% (see Appendix A). Our R functions combining constraint-based search with multiple imputation can be obtained from `https://github.com/bips-hb/micd`.

The score-based algorithm optimized the BIC assuming a multivariate Gaussian distribution. In our analysis, we used the metaheuristic tabu search as implemented in the R package `bnlearn` (Scutari, 2010). This is a variant of hill climbing that uses an adaptive memory to avoid becoming stuck in local optima by carrying on the search after a local optimum has been reached, imposing some tabu on previous solutions (Glover *et al.*, 2007; Scutari *et al.*, 2019). As the output is a DAG we convert it into the corresponding CPDAG, representing the equivalence class, for fair comparison with the outputs of the constraint-based algorithms.

3.2.3.   *Bootstrap*

Robustness of the selected graphs was assessed by using 200 non-parametric bootstrap replications. In each bootstrap replication, $M = 100$ data sets were imputed.

Average Hamming and structural Hamming distances were computed as measures of robustness of the originally selected graph or equivalence class. The Hamming distance counts the number of different edges between two skeletons, whereas the structural Hamming distance counts the number of edge insertions, deletions and flips required to move from one graph to the other. Thus, large values indicate dissimilarity whereas small values indicate similarity of two graphs. Moreover, we compared the bootstrapped graph structures also by recall, precision and false positive rate measures where the original graph was used as truth. For this, we transformed the FCI output into a CPDAG by reducing ∘→ into → and ∘–∘ into −.

4.   Results

First we report the result that was obtained by the constraint-based algorithm. Subsequently we compare this with the alternative approaches and present our sensitivity analyses. In this section we interpret the graphs in terms only of direct or indirect associations found, whereas in Section 5 we discuss the causal interpretation.

Because of the small size of the HPV positive subsample ($n = 24$), the results of our subanalysis must be interpreted with caution. These are reported for completeness in Appendix A.

4.1.   *Constraint-based algorithms*

The CPDAG that was obtained by PC-MI is given in Fig. 1(a) and the PAG that was obtained by FCI-MI in Fig. 1(b). The two methods necessarily yield the same skeleton, i.e. they find the same non-edges, but they can differ on the type of edges as they do here. The graphs are connected and they show five conditional independences indicating absences of direct associations. We note that the PC-MI result is in fact not a CPDAG as this would require some v-structures to support the oriented edges towards HMGA2 and MDM2. The reason is that the algorithm wants v-structures at both nodes, the HMGA2 node and the MDM2 node; hence it cannot orient the edge between HMGA2 and MDM2 without conflict. The FCI-MI algorithm can resolve this conflict by assuming a latent node affecting both HMGA2 and MDM2.

Specifically with regard to the role of HMGA2, this is found to be conditionally independent of BAX and CDKN2A/p14 given CDKN1A/p21, i.e. there is no evidence for a direct association or causal link with these. According to PC-MI, MDM2 and CDKN1A/p21 have direct associations with HMGA2. The additional information that is obtained with FC-MI is that the empirical partial association structure can be explained by assuming a latent variable between HMGA2 and MDM2. Both approaches agree on arrowheads at MDM2 from BAX and at HMGA2 from p21; among others this indicates that HMGA2 and BAX were found to be conditionally independent given CDKN1A/p21, but not when also conditioning on MDM2, and by symmetry for p21 and MDM2 given BAX and HMGA2.

All partial correlations between pairs of variables given a separating set of the other variables are shown in Table 2. Here we find, for instance, that the partial correlation between HMGA2 and MDM2 is $\hat{\rho} = -0.223$.

4.2.   *Alternative analyses*

The DAG that was obtained by the SEM is shown in Fig. 1(c). The skeleton is similar to that of PC and FCI, but with one more edge from CDKN2A/p14 to HMGA2 creating a v-structure. The
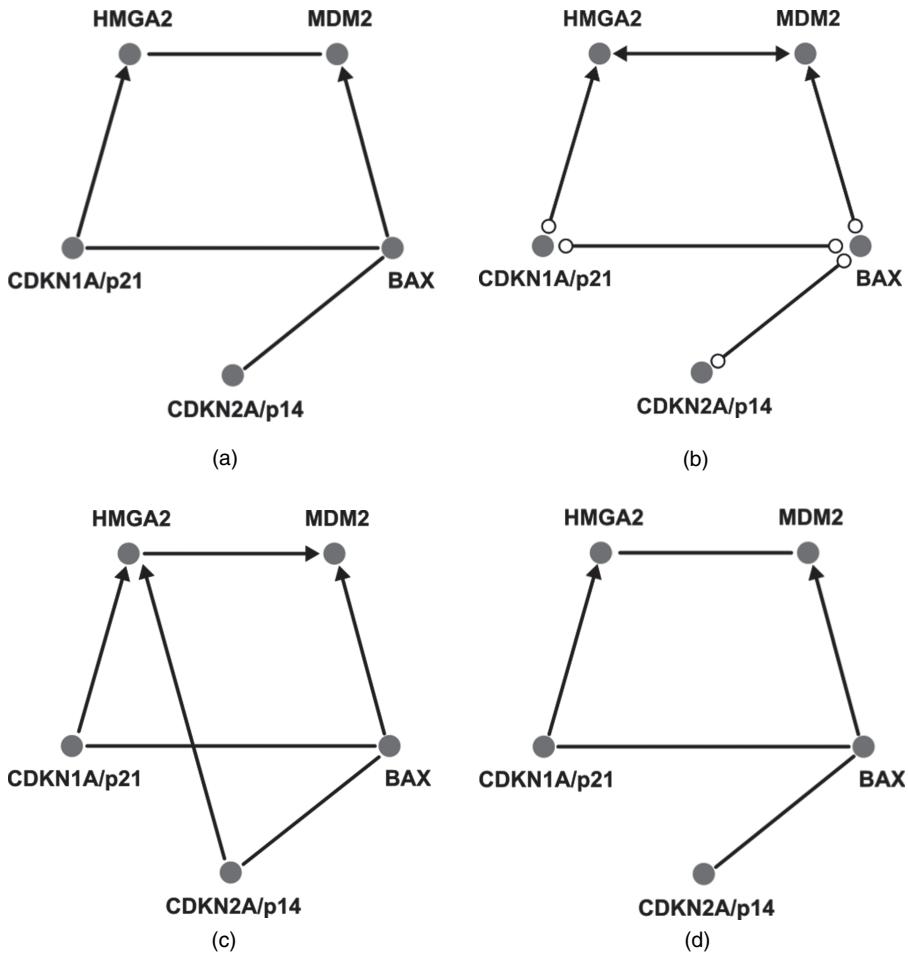
Fig. 1. Selected graphs of 208 HNSCC tumours including HMGA2 (PC stable for CC analysis is based on $N = 159$; for PC stable, PC-MI and FCI-MI, $\alpha = 0.05$; see Section 2.1.2 for interpretation of the various edge types): (a) PC-MI; (b) FCI-MI; (c) SEM; (d) PC stable (CC and TD).

additional edge seems supported by the partial correlation of $\hat{\rho} = -0.218$ between CDKN2A/p14 and HMGA2. A further v-structure results at MDM2. The graph in Fig. 1(d) shows the result from applying PC stable to the CCs only and with TD. We see that in this case the results are the same as for PC-MI. The result for the subsample of HPV positive patients in Fig. 2 shows quite a different association structure, where MDM2 is marginally independent of all other nodes, for instance.

4.3. Stability of results

Note that the CC analysis results in one fewer edge when $\alpha = 0.01$ and one more when $\alpha = 0.1$. The TD analysis finds the same additional edge when $\alpha = 0.1$ but selects the same graphs as found by PC-MI and FCI-MI when $\alpha = 0.01$. Overall, these results do not suggest a serious bias in the CC analysis, but in this data example exploiting more data by using multiple imputation appears slightly more stable.

Table 2. Correlations and partial correlations between two genes adjusted for all relevant other genes as in the original graph†

Gene pair	Conditioned on	Partial correlation	Correlation
CDKN1A/p21, BAX	CDKN2A/p14	0.418	0.480
CDKN2A/p14, BAX	CDKN1A/p21	0.320	0.403
HMGA2, CDKN1A/p21	CDKN2A/p14, BAX	0.293	0.269
MDM2, BAX	CDKN2A/p14, CDKN1A/p21	0.256	0.315
HMGA2, MDM2	CDKN2A/p14, BAX, CDKN1A/p21	−0.223	−0.208
HMGA2, CDKN2A/p14	BAX, CDKN1A/p21	−0.218	−0.135
MDM2, CDKN2A/p14	BAX, CDKN1A/p21	0.114	0.221
CDKN1A/p21, CDKN2A/p14	BAX	0.107	0.279
MDM2, CDKN1A/p21	CDKN2A/p14, BAX	−0.063	0.109
HMGA2, BAX	CDKN1A/p14, CDKN1A/p21	0.007	0.074

†Edges selected by PC-MI or FCI-MI are marked in italics.
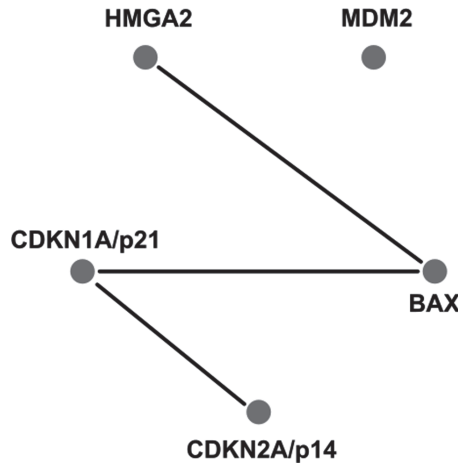
Fig. 2. Selected CPDAG for HPV positive patients ($n = 24$): all algorithms selected the same CPDAG (PC-MI and FCI-MI; $\alpha = 10\%$)

Table 3. Hamming and structural Hamming distances for bootstrapped graphs†

Algorithm	Hamming distance				Structural Hamming distance			
	Mean	SD	Median	IQR	Mean	SD	Median	IQR
PC-MI	1.11	1.10	1	2	3.13	1.66	3	3
FCI-MI	1.15	1.09	1	2	3.80	1.67	4	2
SEM	0.98	1.08	1	2	3.23	2.27	4	5

†Smaller values are better. IQR, interquartile range; PC-MI and FCI-MI with $\alpha = 0.05$.

The bootstrap analysis showed for all three algorithms similar Hamming distances with about one edge deviation between two selected skeletons in the median (Table 3). The structural Hamming distances were also comparable between approaches (see Table 3). Fig. 3 shows the boxplots of the performance measures (recall, false positive rate and precision); the constraint-based algorithms slightly outperform the score-based SEM regarding stability.
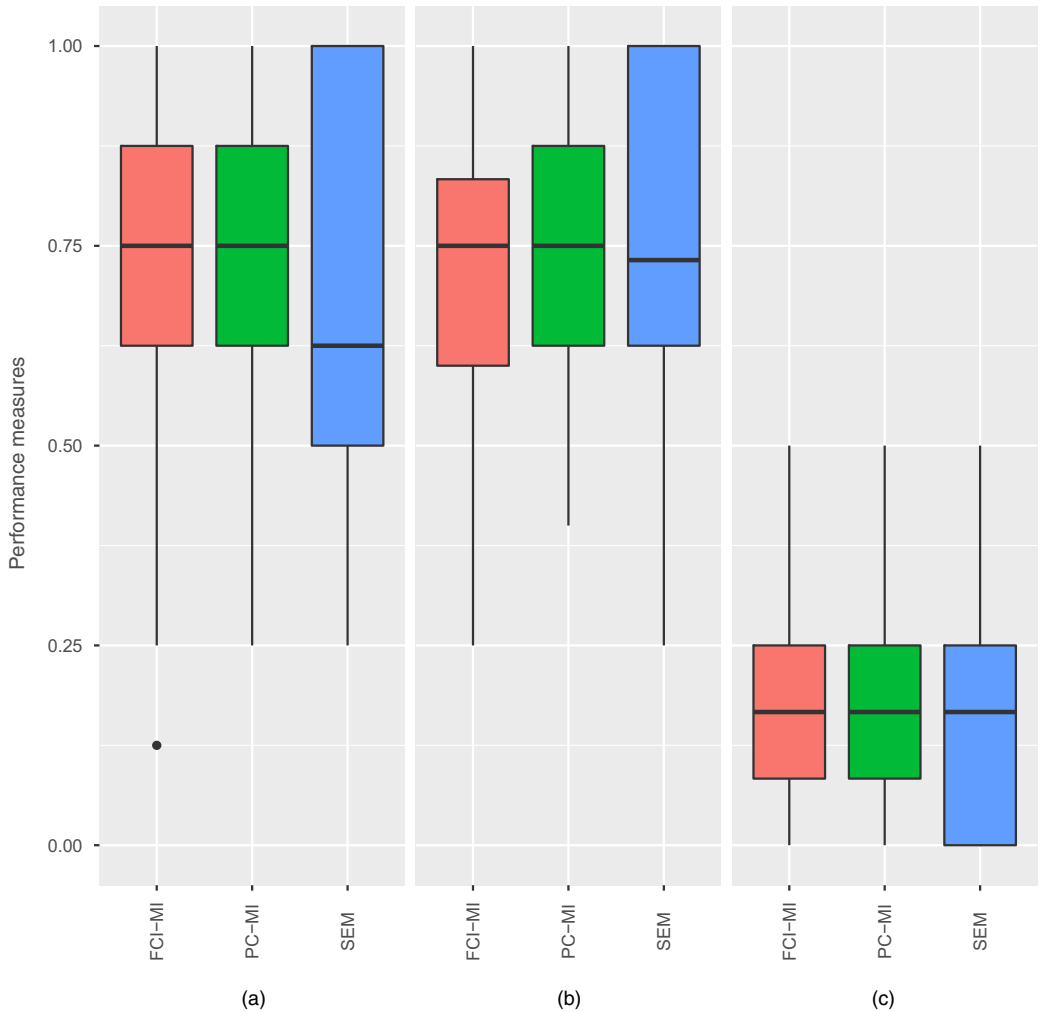
Fig. 3. (a) Recall, (b) precision and (c) false positive rate boxplots of the bootstrapped graphs for each causal discovery algorithm (PC-MI and FCI-MI; $\alpha = 5\%$)

Table 4. Relative frequencies of edges between two genes being present in the graphs obtained from 200 non-parametric bootstrap samples for PC-MI/FCI-MI/SEM†

Gene	CDKN2A/p14	BAX	HMGA2	MDM2	CDKN1A/p21
CDKN2A/p14	—	*77/85/63*	20/18 /41	8/9/6	26/32/20
BAX	*88/94/87*	—	0/0/0	*85/86/81*	*89/86/78*
HMGA2	24 /22/26	0/0/1	—	*78/72/72*	*86/92/48*
MDM2	8/10/8	*48/78/32*	*69/64/28*	—	0/0/1
CDKN1A/p21	6/32/14	*88/96/78*	*94/94/66*	0/0/0	—

†Edge directions are read from row to column. Entries in italics indicate those edges selected in the original graphs. PC-MI and FCI-MI with $\alpha = 0.05$. PC-MI and SEM: an edge from for example MDM2 to HMGA2 was counted if either MDM2 $\rightarrow$ or $-$ HMGA2 was selected (69%, 28%); FCI-MI with an edge from for example MDM2 to HMGA2 was counted if MDM2 $\rightarrow$, $-\circ$ or $\leftarrow\circ$ HMGA2 was selected (64%).

Table 4 enables us to assess the stability of individual (non-)edges. Generally we see again a little more variability for the score-based algorithm than for the constraint-based algorithms. The absence of edges between MDM2 and CDKN1A/p21 as well as between HMGA2 and BAX were very stable as these are never selected, whereas the absence of an edge between MDM2 and CDKN2A/p14 was a little less but still quite stable. The most stable edges were between CDKN1A/p21 and BAX and from BAX to MDM2; further, although the edge from CDKN1A/p21 to HMGA2 was quite stable, the reverse direction occurred almost with equal frequency, so we cannot be very confident in the orientation of this edge. Clearly there was also some uncertainty around the edge between HMGA2 and MDM2, both regarding its orientation as well as (but to a lesser extent) its presence.

5. Discussion

5.1. Causal interpretation of results

As set out in Section 2.2, under additional assumptions the outputs of the various algorithms can be given a causal interpretation. Under causal sufficiency, the result of the PC algorithm suggested that there is evidence that CDKN1A/p21 is a direct cause of HMGA2 whereas BAX is a direct cause of MDM2; but the causal direction between HMGA2 and MDM2, or between the other pairs linked by undirected edges, could not be decided. The result also suggested that BAX and CDKN2A/p14 have no direct causal effect on HMGA2, but the data were compatible with indirect causal effects; similarly, CDKN1A/p21 and CDKN2A/p14 had no direct causal effect on MDM2, but there could again be indirect causal effects.

Although the results of PC-MI are interesting, the assumption of causal sufficiency is not plausible in our analysis: there are more than 70 proteins binding to p53 (Inoue *et al.*, 2016) and unobserved genes and biomarkers known to be relevant to the p53 pathway. Hence, for a more realistic causal interpretation, we should focus on the FCI-MI result. This interpretation is quite different: relaxing the assumption of causal sufficiency leads in this example to the result that all direct associations (edges) that are detected by PC-MI could be due to latent confounding. Two aspects were still interesting and relevant.

(a) The FCI output suggested that HMGA2 is not causal for CDKN1A/p21, but, vice versa, the latter could be causal for HMGA2. However, as mentioned earlier, the bootstrap results (Table 6 in Appendix A.1) suggested that the direction of this particular edge was very unstable. This finding, the direct association with a causal relationship between HMGA2 and p21, was consistent with prior research: Narita *et al.* (2006) suggested that HMGA proteins contribute to a stable state of the cell known as senescence induced by prosenescence signals. Increased gene expression values of the molecular senescence marker p21 might be such a signal.

(b) Because of the absence of any outgoing edges from HMGA2, no DAG in the equivalence class is compatible with HMGA2 being causal for any of the other gene expressions. This means, that, on the basis of our analysis, we see no clear support for the hypothesis that HMGA2 is itself a key regulator gene in the p53 pathway and that it would be useful as a therapeutic target.

5.2. Human papilloma-virus positive subsample analysis

According to our subsample analysis, the gene regulatory network had a different shape for HPV positive HNSCC (see Fig. 2). Here, HGMA2 had a direct association with BAX. As BAX is a proapoptic protein involved in the programmed self-destruction of cells with damaged

DNA, this is compatible with prior research: Shi *et al.* (2015) demonstrated that HMGA2 could induce apoptosis in primary human cells. Hence, our results were consistent with prior research on how the regulatory structure differs for HPV positive HNSCC. Because of the small subsample size, these results can only be indicative and need to be confirmed in larger studies.

5.3.   *Strengths and limitations of application*

Our application of causal discovery to the HNSCC data relied on a representative, clinically and epidemiologically very well-characterized patients' collective. The tumour-tissue-based gene expression values were analysed by using established laboratory analysis methods. The sample size was substantial compared with other studies focusing on HMGA2 in human cancer (Huang *et al.*, 2018). Given the sample size and the low number of genes included, it is not surprising that the results that we obtained were reasonably stable across different methods, as well as regarding the choices of tuning parameters and with a view to variability assessed by the non-parametric bootstrap.

However, the limited number of investigated genes of the p53 signalling pathway and missing information on TP53 mutation status also constitute a drawback of the study. Although the FCI algorithm allows for latent variables, our results illustrate the price for greater generality when relaxing the causal sufficiency assumption: all of the direct associations that were found could be due to unobserved confounding. Hence, as is to be expected, the output of an FCI algorithm will typically be much more vague than when causal sufficiency can be assumed.

In view of the above limitations it is interesting to investigate the replicability of our results on a different data set with a more extensive gene set constituting the p53 pathway. We therefore carried out additional analyses with data on 73 genes obtained from *The Genome Atlas of Cancer Head–Neck Squamous Cell Carcinoma* (Cancer Genome Atlas Network (2015); see Appendix A.3). The results did not clearly support nor contradict our primary analysis. Importantly, HMGA2 was again not found to be causal for any of the other four gene expressions and in the extended gene set it was causal only for THBS1. Even in the extended gene set, the FCI algorithm found many associations to be due to latent confounding (50 of 167 edges were bidirected). However, the results based on the data of the Cancer Genome Atlas Network (2015) should be taken with a pinch of salt: these data contain no missing values and are probably CCs only with incomplete cases omitted; moreover, as far as we can tell from the available documentation, there are major differences in the study design and measurement methods. It is unclear whether the gene expression dependence structure can be expected to be stable across such differences between data sets. A more fundamental question may be the suitability of causal discovery for dynamic gene regulatory processes. All causal discovery methods implicitly rely on the assumption that the causal structure can be appropriately represented by a DAG (Dawid, 2010). This may, for instance, not be so if the measurements taken are only a snapshot of a time-varying system especially if it also exhibits feedback (Aalen *et al.*, 2016). Gene regulation is such a dynamic process; hence it might be more appropriate to use longitudinal data and methods that are suitable for dynamic networks (Husmeier, 2006). Such data are currently not available for the p53 pathway and HNSCC patients.

5.4.   *Strengths and limitations of methods*

We have proposed and applied a first suggestion for combining multiple imputation with constraint-based causal discovery. The key idea was to pool the conditional independence test

across the imputed data sets. In the application we found that our procedure gave plausible and stable results across our sensitivity analyses. These were in fact slightly more stable than for the score-based approach using the EM algorithm. The method is straightforward to implement (see R package `micd`; `https://github.com/bips-hb/micd`).

However, the current implementations of PC-MI and FCI-MI can handle only continuous variables, but the principle is entirely general and the software will be extended in future work. SEMs, as implemented in `bnlearn`, can be applied to mixed discrete and continuous variables. Regarding the handling of missing data, PC-MI and FCI-MI are more flexible than SEMs as they allow for very general imputation models. In particular, we found it useful that additional variables to those representing the nodes in the graph could be included in the imputation models, strengthening the plausibility of MAR, e.g. hospital or tumour stage. This is not possible with CC and TD analysis, or with the SEM algorithm. If MAR is violated, the graphs selected may be very different and the causal conclusions are then likely to be wrong. Recent advances address causal discovery with missing data (Mohan *et al.*, 2013), especially in combination with TD under missingness not at random (Strobl *et al.*, 2018; Tu *et al.*, 2019). An advantage of using multiple imputation over TD, if MAR holds, is that all the data are used.

Given that the combination of constraint-based algorithms with multiple imputation presents a very promising and flexible approach to causal discovery with incomplete data, future work should establish formal properties, e.g. general conditions on the imputation models and test procedures ensuring consistent estimation of the causal structure.

Acknowledgements

Appendix A

A.1. *Sensitivity analyses*
In sensitivity analyses we applied PC-MI and FCI-MI with α-values 1% and 10%. The selected graphs are presented in Fig. 4, (structural) Hamming distances are shown in Table 5, edge stabilities in Table 6 and boxplots of the classification measures recall, false positive rate and precision in Fig. 5.

A.2. *Human papilloma-virus positive subgroup analysis*
It is well known that HPV-induced HNSCC tumours have pervasive distinct differences compared with HPV negative HNSCC. We therefore also investigated the graph structure in the subgroup of $n = 24$ patients
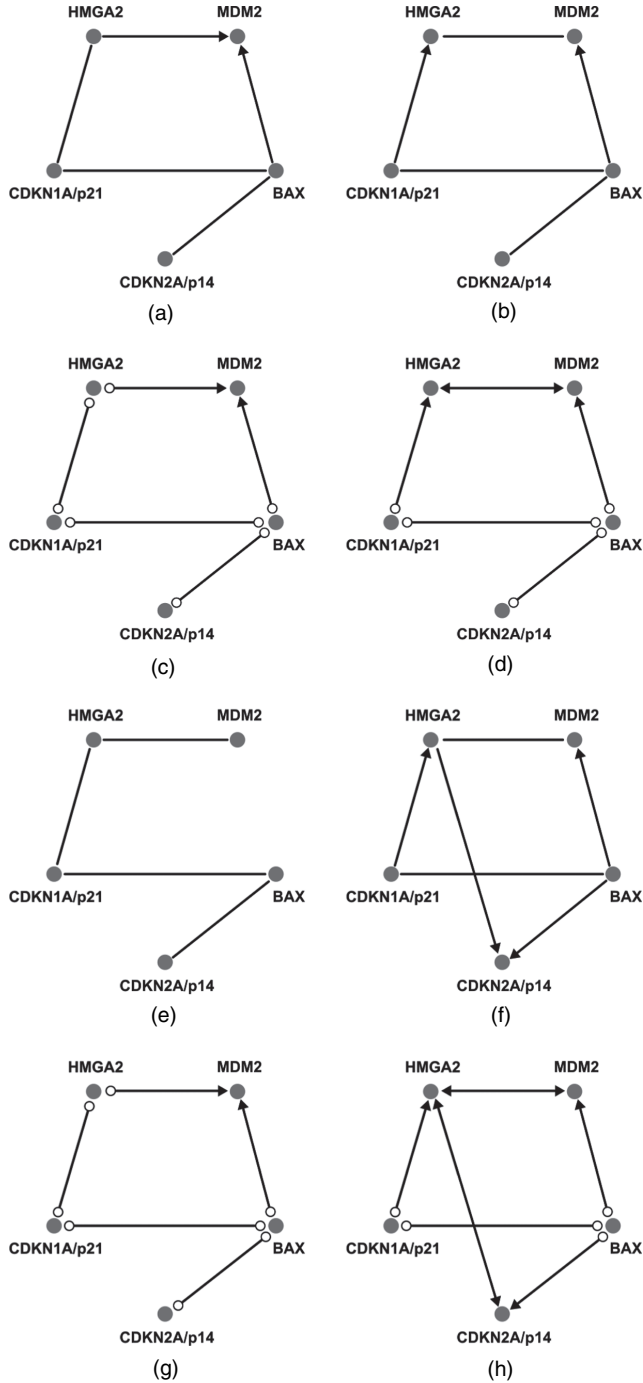
Fig. 4. Selected graphs for 208 HNSCC tumours including HMGA2 (PC stable (CC) is based on CC analysis ($N = 159$); see Section 2.1.2 for an interpretation of different edge types): (a) PC-MI, $\alpha = 1\%$; (b) PC-MI, $\alpha = 10\%$; (c) FCI-MI, $\alpha = 1\%$; (d) FCI-MI, $\alpha = 10\%$; (e) PC-stable (CC), $\alpha = 1\%$; (f) PC-stable (CC), $\alpha = 10\%$; (g) FCI-stable (TD), $\alpha = 1\%$; (h) FCI-stable (TD), $\alpha = 10\%$

being positive for HPV-16 or HPV-18 subtypes. Because of this small sample size, we considered only the nominal significance level of α to 10%. Most HPV positive patients have overexpressed CDKN2A/p14 values (87.5%).

Fig. 2 shows the graph structure that was found by both PCI-MI and FCI-MI. All discovery algorithms selected the same undirected graph, which allows no causal interpretation other than that MDM2 appears to be non-causal for any of the other genes. The HPV positive graph shares only the edge between CDKN1A/p21 and BAX with the graphs in Fig. 1 and includes the additional edge between HMGA2 and BAX. Table 7 shows the stability of individual (non-)edges.

A.3. The Cancer Genome Atlas

We compared our results with external data by analysing the publicly available transcriptome profiling data of HNSCC patients provided by *The Cancer Genome Atlas Head-Neck Squamous Cell Carcinoma* (Cancer Genome Atlas Network, 2015) and downloaded with the R package `TCGAbiolinks` (Colaprico *et al.*, 2016).

The gene expression values are generated counts of the reads mapped to each gene from an RNA sequencing alignment and normalized by fragments per kilobase of transcript per million mapped reads

Table 5. Hamming and structural Hamming distances for bootstrapped graphs†

Algorithm	α (%)	Hamming distance				Structural Hamming distance			
		Mean	SD	Median	IQR	Mean	SD	Median	IQR
PC-MI	1	1.32	1.00	1	1	2.56	1.28	2	1
	10	1.22	0.97	1	1.75	3.59	1.65	4	3
FCI-MI	1	1.28	1.10	1	2	2.74	1.50	2	2
	10	1.17	0.90	1	1	4.08	1.69	4	2
SEM	—	0.98	1.08	1	2	3.23	2.27	4	5

†Smaller values are better. IQR, interquartile range; α, significance value.

Table 6. Relative frequencies of edges between two genes being present in the graphs obtained from 200 non-parametric bootstrap samples for PC-MI/FCI-MI/SEM†

Gene	α (%)	Results for the following genes:				
		CDKN2A/p14	BAX	HMGA2	MDM2	CDKN1A/p21
CDKN2A/p14	1	—	78/82/30	10/11/41	6/5/4	9/12/18
	10	—	65/84/63	26/26/41	14/9/6	34/40/20
BAX	1	84/83/69	—	0/0/0	68/74/74	97/94/55
	10	88/96/87	—	0/0/0	96/91/81	82/76/78
HMGA2	1	12/10/24	0/0/1	—	46/52/66	78/88/35
	10	38/34/26	0/0/1	—	94/81/72	83/92/48
MDM2	1	6/6/6	48/72/22	40/52/24	—	0/0/1
	10	14/14/8	58/75/32	82/71/28	—	0/0/1
CDKN1A/p21	1	4/12/14	96/100/45	86/88/62	0/0/0	—
	10	14/40/14	86/95/78	92/94/66	0/0/0	—

†Edge directions are read from row to column. Entries in italics indicate those edges selected in the original graphs. PC-MI (1%), SEM, an edge from HMGA2 to MDM2 was counted if either HMGA2 → or − MDM2 was selected (43%, 72%); FCI-MI (1%), an edge from for example HMGA2 to MDM2 was counted if HMGA2 →, −∘ or ←∘ MDM2 (48%) was selected; α, significance value.
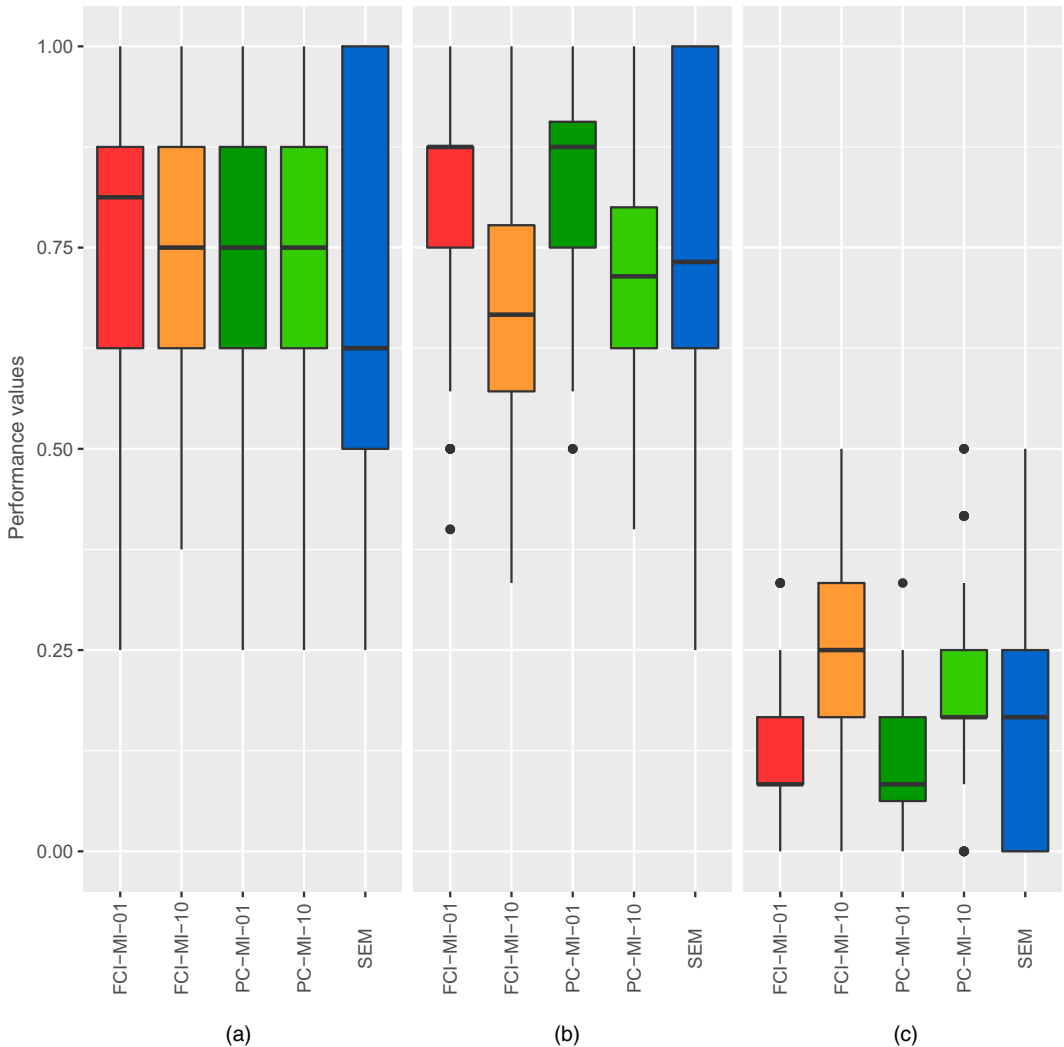
Fig. 5. (a) Recall, (b) precision and (c) false positive rate boxplots of the bootstrapped graphs for each causal discovery algorithm (PC-MI, FCI-MI; $\alpha = 1\%$, 10%)

calculation; see `https://docs.gdc.cancer.gov/Data/Bioinformatics_Pipelines/Expression_mRNA_Pipeline`). The original data set contains $N = 546$ cases of which we included only Caucasians with the same diagnoses as in our study data. This resulted in $N = 392$ observations. The data of the Cancer Genome Atlas Network (2015) differ in three important ways from our study.

First, the Cancer Genome Atlas Network (2015) uses an untargeted RNA sequencing strategy and quantified messenger RNA by HT-Seq raw read counts whereas in our study realtime quantitative PCR with specially designed primers for the selected genes were applied which allow detecting precise upregulation and dynamic range of targeted genes. Second, gene CDKN2A encodes two different proteins, p16 and p14ARF. In the data of the Cancer Genome Atlas Network (2015) alternative splice products cannot be distinguished whereas in our study only the alternative reading frame p14ARF was measured by using a special primer. Third, the Genomic Data Commons Data Portal provides only cases with complete data and therefore the data cannot serve as an application for our proposed multiple-imputation method. Both data sets are hence not concordant and might not be comparable regarding the biological research question.

Table 7. Relative frequencies of edges between two genes being present in the graphs obtained from 200 non-parametric bootstrap samples for PC-MI/FCI-MI/SEM in the subgroup of HPV positive patients†

Gene	Results for the following genes:				
	CDKN2A/p14	*BAX*	*HMGA2*	*MDM2*	*CDKN1A/p21*
CDKN2A/p14	—	17/21/22	14/6/16	10/11/9	77/88/64
BAX	16/21/17	—	*55/57/58*	1/1/4	*53/47/64*
HMGA2	16/6/17	55/57/67	—	3/3/6	13/14/20
MDM2	12/11/40	1/1/8	4/3/11	—	1/1/5
CDKN1A/p21	*80/88/77*	*51/47/66*	14/14/21	1/1/0	—

†Edge directions are read from row to column; for example an edge from HMGA2 to MDM2 was counted if PC-MI or the SEM selected HMGA2 $\rightarrow$ or $-$ MDM2 (2%, 6%); if FCI-MI selected HMGA2 $\rightarrow$, $\multimap$ or $\leftarrow\!\!\circ$ MDM2 (3%). Entries in italics indicate those edges selected in the original graphs. PC-MI and FCI-MI; $\alpha = 10\%$.
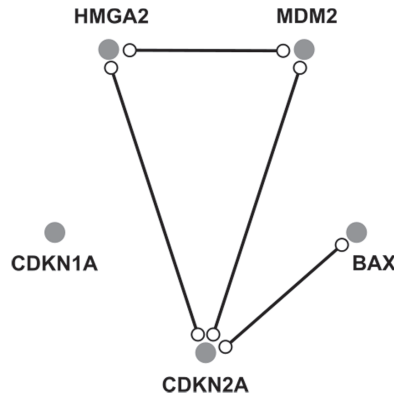

Fig. 6. Selected graph by using a subset of the TCGA data ($N = 392$) applying FCI stable ($\alpha = 5\%$)

Table 8. Relative frequencies of edges between two genes being present in the graphs based on the subset of the five genes (5g)/genes included in the p53 pathway (p53) obtained from 200 non-parametric bootstrap samples†

Gene	Results for the following genes:				
	CDKN2A, 5g/p53	*BAX, 5g/p53*	*HMGA2, 5g/p53*	*MDM2, 5g/p53*	*CDKN1A, 5g/p53*
CDKN2A	0/0	100/5	74/0	96.5/1.5	0.5/0
BAX	96.5/7.5	0/0	0/0	0.5/0	11/0
HMGA2	74/0	0/0	0/0	47/7	9.5/0
MDM2	97/1	0.5/0	47/7	0/0	7.5/0.5
CDKN1A	0.5/0	11/0	9.5/0	7.5/0.5	0/0

†FCI was used for model selection ($\alpha = 5\%$). Edge directions are read from row to column. Entries in italics indicate those edges selected in the original graphs. An edge from for example HMGA2 to MDMD2 was counted if HMGA2 $\rightarrow$, $\multimap$ or $\leftarrow\!\!\circ$ MDM2 was selected (47% and 7% respectively).
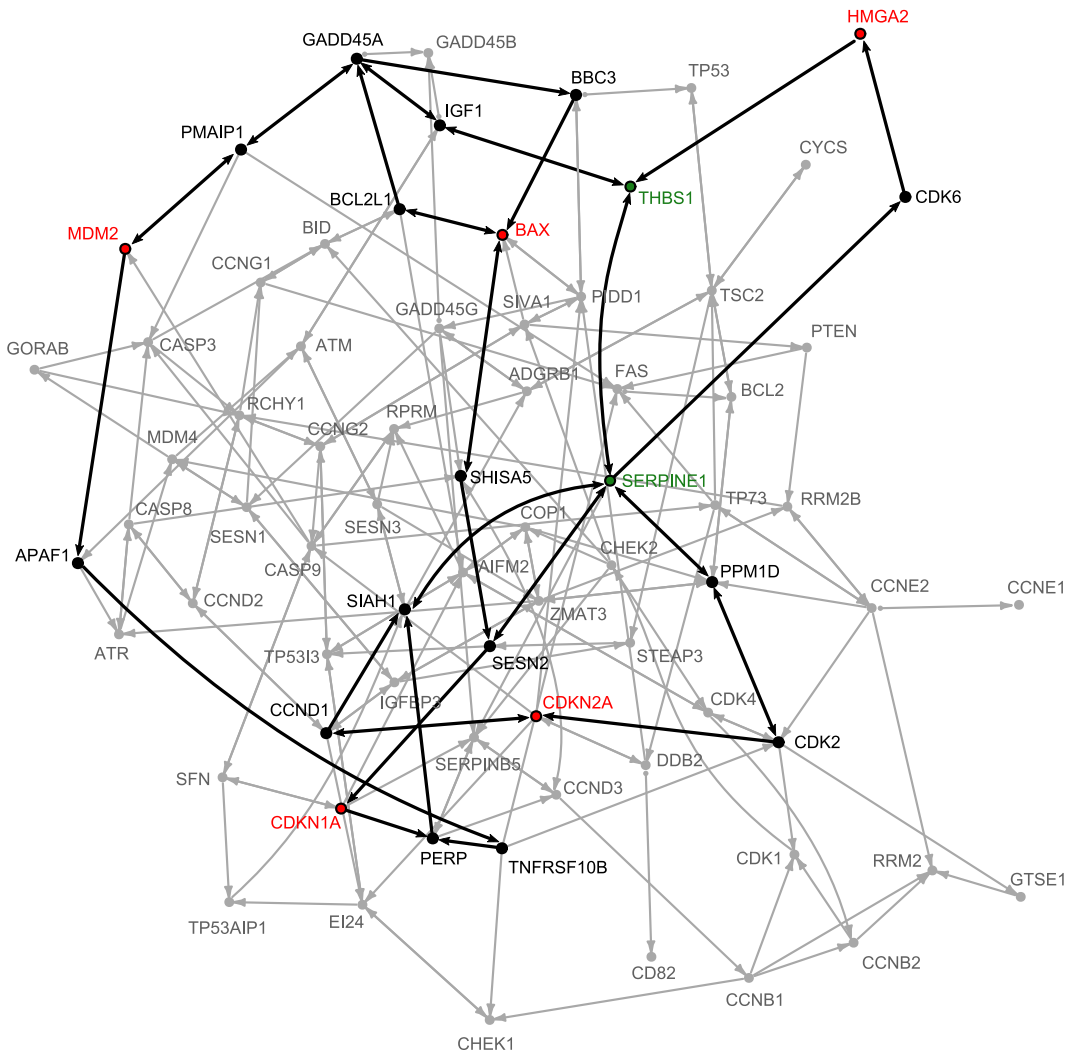
Fig. 7. Selected graph of genes included in the p53 pathway for $N = 392$ HNSCC tumours from the data set of the Cancer Genome Atlas Network (2015) using PC stable ($\alpha = 5\%$); for readability, only the subgraph of the five genes of interest and all 30 genes lying on the shortest path between them which are contained in the p53 pathway are shown; the shortest paths to and from HGMA2 are highlighted; the gene expression of HMGA2 is conditional independent of BAX, MDM2, CDKN1A and CDKN2A (marked in red) given the gene expressions of SERPINE1 and THBS1 (marked in green)

We applied FCI stable as in Section 2 to perform causal discovery on the five genes HMGA2, MDM2, BAX, CDKN1A and CDKN2A (Fig. 6) as well as on all 73 genes contained in the p53 pathway according to the *Kyoto Encyclopedia of Genes and Genomes* (Ogata and Goto (2000); Fig. 7).

Fig. 6 suggests that all four edges could go either direction or be due to latent confounding so that the presence and directions of any causal relationships cannot be determined. The graph structure agrees on two adjacencies and three non-adjacencies but is otherwise different from Fig. 1. Here, CDKN1A is non-causal for HMGA2 and CDKN2A is the most central node in the graph on the basis of its connectivity. This difference compared with our primary analysis could be a consequence of the different splicing products. The graph on the extended gene set in Fig. 6 suggests that HMGA2 has two important neighbours: THBS1 and SERPINE1 separating HMGA2 from the other genes in the p53 pathway.

As the only directed path from HMGA2 is identical with the edge to THBS1, we find no clear support for a central causal role of HMGA2 in the p53 pathway. Moreover, the edge between HMGA2 and MDM2 is subject to some uncertainty (47% in the bootstrap replications for the graph in Fig. 6; see Table 8).

References

Aalen, O. O., Røysland, K., Gran, J. M., Kouyos, R. and Lange, T. (2016) Can we believe the DAGs?: A comment on the relationship between causal DAGs and mechanisms. *Statist. Meth. Med. Res.*, **25**, 2294–2314.

Albieri, V. and Didelez, V. (2014) Comparison of statistical methods for finding network motifs. *Statist. Appl. Genet. Molec. Biol.*, **13**, 403–422.

Bansal, M., Belcastro, V., Ambesi-Impiombato, A. and di Bernardo, D. (2007) How to infer gene networks from expression profiles. *Molec. Syst. Biol.*, **3**, article 78.

van Buuren, S. (2018) *Flexible Imputation of Missing Data*. Boca Raton: Chapman and Hall–CRC.

van Buuren, S. and Groothuis-Oudshoorn, K. (2011) mice: multivariate imputation by chained equations in R. *J. Statist. Softwr.*, **45**, 1–67.

Cancer Genome Atlas Network (2015) Comprehensive genomic characterization of head and neck squamous cell carcinomas. *Nature*, **517**, 576.

Chickering, D. M. (1995) A transformational characterization of equivalent Bayesian network structures. In *Proc. 11th Conf. Uncertainty in Artificial Intelligence* (eds P. Besnard and S. Hanks), pp. 87–98. San Francisco: Morgan Kaufmann.

Chickering, D. M. (2002) Learning equivalence classes of Bayesian-network structures. *J. Mach. Learn. Res.*, **2**, 445–498.

Chu, T., Glymour, C., Scheines, R. and Spirtes, P. (2003) A statistical problem for inference to regulatory structure from associations of gene expression measurements with microarrays. *Bioinformatics*, **19**, 1147–1152.

Colaprico, A., Silva, T. C., Olsen, C., Garofano, L., Cava, C., Garolini, D., Sabedot, T., Malta, T. M., Pagnotta, S. M., Castiglioni, I., Ceccarelli, M., Bontempi, G. and Noushmehr, H. (2016) TCGAbiolinks: an R/Bioconductor package for integrative analysis of tcga data. *Nucleic Acids Res.*, **44**, article e71.

Colombo, D. and Maathuis, M. H. (2014) Order-independent constraint-based causal structure learning. *J. Mach. Learn. Res.*, **15**, 3741–3782.

Colombo, D., Maathuis, M. H., Kalisch, M. and Richardson, T. S. (2012) Learning high-dimensional directed acyclic graphs with latent and selection variables. *Ann. Statist.*, **40**, 294–321.

Cox, D. R. and Wermuth, N. (1996) *Multivariate Dependencies: Models, Analysis and Interpretation*. Boca Raton: Chapman and Hall–CRC.

van Dam, S., Vosa, U., van der Graaf, A., Franke, L. and de Magalhaes, J. P. (2018) Gene co-expression analysis for functional classification and gene-disease predictions. *Brief. Bioinform.*, **19**, 575–592.

D'Angelo, G. M., Luo, J. and Xiong, C. (2012) Missing data methods for partial correlations. *J. Biometr. Biostatist.*, **3**, 1–7.

Dawid, A. P. (2010) Beware of the DAG! *J Mach. Learn. Res.*, **6**, 59–86.

Dempster, A. P., Laird, N. M. and Rubin, D. B. (1977) Maximum likelihood from incomplete data via the EM algorithm. *J. R. Statist. Soc.* B, **39**, 1–38.

Didelez, V. (2018) Causal concepts and graphical models. In *Handbook of Graphical Models* (eds M. Maathuis, M. Drton, S. L. Lauritzen and M. Wainwright), 1st edn. ch. 15. Boca Raton: CRC Press.

Didelez, V. and Pigeot, I. (1998) Maximum likelihood estimation in graphical models with missing values. *Biometrika*, **85**, 960–966.

D'Souza, G., Anantharaman, D., Gheit, T., Abedi-Ardekani, B., Beachler, D. C., Conway, D. I., Olshan, A. F., Wunsch-Filho, V., Toporcov, T. N., Ahrens, W., Wisniewski, K., Merletti, F., Boccia, S., Tajara, E. H., Zevallos, J. P., Levi, J. E., Weissler, M. C., Wright, S., Scelo, G., Mazul, A. L., Tommasino, M., Brennan, P. and Cadoni, G. (2016) Effect of HPV on head and neck cancer patient survival, by region and tumor site: a comparison of 1362 cases across three continents. *Oral Oncol.*, **62**, 20–27.

Faraji, F., Schubert, A. D., Kagohara, L. T., Tan, M., Xu, Y., Zaidi, M., Fortin, J.-P., Fakhry, C., Izumchenko, E., Gaykalova, D. A. and Fertig, E. J. (2018) The genome-wide molecular landscape of HPV-driven and HPV-negative head and neck squamous cell carcinoma. In *Molecular Determinants of Head and Neck Cancer* (eds B. Burtness and E. A. Golemis), 2nd edn. pp. 293–325. Berlin: Springer.

Fisher, R. A. (1924) The distribution of the partial correlation coefficient. *Metron*, **3**, 329–332.

Friedman, N. (1997) Learning belief networks in the presence of missing values and hidden variables. In *Proc. 14th Int. Conf. Machine Learning*, pp. 125–133.

Friedman, N., Goldszmidt, M. and Wyner, A. (1999) Data analysis with Bayesian networks: a bootstrap approach. In *Proc. 15th Conf. Uncertainty in Artificial Intelligence* (eds K. B. Laskey and H. M. Prade), pp. 196–205. San Francisco: Morgan Kaufmann.

Friemel, J., Foraita, R., Günther, K., Heibeck, M., Günther, F., Pflueger, M., Pohlabeln, H., Behrens, T., Bullerdiek, J., Nimzyk, R. and Ahrens, W. (2016) Pretreatment oral hygiene habits and survival of head and neck squamous cell carcinoma (HNSCC) patients. *BMC Oral Hlth*, **16**, article 33.

Gavathiotis, E., Reyna, D. E., Bellairs, J. A., Leshchiner, E. S. and Walensky, L. D. (2012) Direct and selective small-molecule activation of proapoptotic BAX. *Nat. Chem. Biol.*, **8**, 639–645.

Gillispie, S. B. and Perlman, M. D. (2002) The size distribution for Markov equivalence classes of acyclic digraph models. *Artif. Intell.*, **141**, 137–155.

Glover, F., Laguna, M. and Marti, R. (2007) Principles of tabu search. In *Handbook of Approximation Algorithms and Metaheuristics* (ed. T. Gonzalez), vol. 23, pp. 1–12. New York: Chapman and Hall.

Hamming, R. W. (1950) Error detecting and error correcting codes. *Bell Syst. Tech. J.*, **29**, 147–160.

Hanahan, D. and Weinberg, R. A. (2011) Hallmarks of cancer: the next generation. *Cell*, **144**, 646–674.

Heinze-Deml, C., Maathuis, M. H. and Meinshausen, N. (2018) Causal structure learning. *A. Rev. Statist. Appl.*, **5**, 371–391.

Hetland, T. E., Holth, A., Kærn, J., Flørenes, V. A., Tropé, C. G. and Davidson, B. (2012) HMGA2 protein expression in ovarian serous carcinoma effusions, primary tumors, and solid metastases. *Virch. Arch.*, **460**, 505–513.

Hotelling, H. (1953) New light on the correlation coefficient and its transforms (with discussion). *J. R. Statist. Soc.* B, **15**, 193–232.

Huang, B., Yang, J., Cheng, Q., Xu, P., Wang, J., Zhang, Z., Fan, W., Wang, P. and Yu, M. (2018) Prognostic value of HMGA2 in human cancers: a meta-analysis based on literatures and TCGA datasets. *Front. Physiol.*, **9**, article 776.

Husmeier, D. (2006) Inferring genetic regulatory networks from microarray experiments with Bayesian networks. In *Probabilistic Modeling in Bioinformatics and Medical Informatic* (eds D. Husmeier, R. Dybowski and S. Roberts), ch. 8, pp. 239–267. London: Springer Science and Business Media.

Husmeier, D., Dybowski, R. and Roberts, S. (2006) *Probabilistic Modeling in Bioinformatics and Medical Informatics*. London: Springer Science and Business Media.

Inoue, K., Fry, E. A. and Frazier, D. P. (2016) Transcription factors that interact with p53 and Mdm2. *Int. J. Cancer*, **138**, 1577–1585.

Ji, Q., Hao, X., Meng, Y., Zhang, M., Desano, J., Fan, D. and Xu, L. (2008) Restoration of tumor suppressor miR-34 inhibits human p53-mutant gastric cancer tumorspheres. *BMC Cancer*, **8**, article 266.

Kalisch, M. and Bühlmann, P. (2007) Estimating high-dimensional directed acyclic graphs with the PC-algorithm. *J. Mach. Learn. Res.*, **8**, 613–636.

Kalisch, M. and Bühlmann, P. (2014) Causal structure learning and inference: a selective review. *Qual. Technol. Quant. Mangmnt*, **11**, 3–21.

Kalisch, M., Mächler, M., Colombo, D., Maathuis, M. H. and Bühlmann, P. (2012) Causal inference using graphical models with the R package pcalg. *J. Statist. Softwr.*, **47**, 1–26.

Klemke, M., Meyer, A., Nezhad, M. H., Bartnitzke, S., Drieschner, N., Frantzen, C., Schmidt, E. H., Belge, G. and Bullerdiek, J. (2009) Overexpression of HMGA2 in uterine leiomyomas points to its general role for the pathogenesis of the disease. *Genes Chromsm. Cancer*, **48**, 171–178.

Lagiou, P., Georgila, C., Minaki, P., Ahrens, W., Pohlabeln, H., Benhamou, S., Bouchardy, C., Slamova, A., Schejbalova, M., Merletti, F., Richiardi, L., Kjaerheim, K., Agudo, A., Castellsague, X., Macfarlane, T. V., Macfarlane, G. J., Talamini, R., Barzan, L., Canova, C., Simonato, L., Lowry, R., Conway, D. I., McKinney, P. A., Znaor, A., McCartan, B. E., Healy, C., Nelis, M., Metspalu, A., Marron, M., Hashibe, M. and Brennan, P. J. (2009) Alcohol-related cancers and genetic susceptibility in Europe: the ARCAGE project: study samples and data collection. *Eur. J. Cancer Prevn*, **18**, 76–84.

Lallemant, B., Evrard, A., Combescure, C., Chapuis, H., Chambon, G., Raynal, C., Reynaud, C., Sabra, O., Joubert, D., Hollande, F., Lallemant, J. G., Lumbroso, S. and Brouillet, J. P. (2009) Reference gene selection for head and neck squamous cell carcinoma gene expression studies. *BMC Molec. Biol.*, **10**, article 78.

Lauritzen, S. L. (1996) *Graphical Models*. Oxford: Clarendon.

Levine, A. J., Hu, W. and Feng, Z. (2006) The P53 pathway: what questions remain to be explored? *Cell Death Differentn*, **13**, 1027–1036.

Little, R. J. A. and Rubin, D. B. (2002) *Statistical Analysis with Missing Data*, 2nd edn. Hoboken: Wiley.

Liu, T.-C., Jin, X., Wang, Y. and Wang, K. (2017) Role of epidermal growth factor receptor in lung cancer and targeted therapies. *Am. J. Cancer Res.*, **7**, 187–202.

Maathuis, M. H., Colombo, D., Kalisch, M. and Bühlmann, P. (2010) Predicting causal effects in large-scale systems from observational data. *Nat. Meth.*, **7**, 247–248.

Maathuis, M. H., Kalisch, M. and Bühlmann, P. (2009) Estimating high-dimensional intervention effects from observational data. *Ann. Statist.*, **37**, 3133–3164.

Maathuis, M. H. and Nandy, P. (2016) A review of some recent advances in causal inference. In *Handbook of Big Data* (eds P. Bühlmann, P. Drineas, M. Kane and M. van der Laan), 1st edn, ch. 21, pp. 387–408. Boca Raton: CRC Press.

Markowski, D. N., von Ahsen, I., Nezhad, M. H., Wosniok, W., Helmke, B. M. and Bullerdiek, J. (2010) HMGA2 and the p19Arf-TP53-CDKN1A axis: a delicate balance in the growth of uterine leiomyomas. *Genes Chromsm. Cancer*, **49**, 661–668.

Markowski, D. N., Helmke, B. M., Belge, G., Nimzyk, R., Bartnitzke, S., Deichert, U. and Bullerdiek, J. (2011) HMGA2 and p14Arf: major roles in cellular senescence of fibroids and therapeutic implications. *Anticancer Res.*, **31**, 753–761.

Millon, R., Muller, D., Schultz, I., Salvi, R., Ghnassia, J. P., Frebourg, T., Wasylyk, B. and Abecassis, J. (2001) Loss of MDM2 expression in human head and neck squamous cell carcinomas and clinical significance. *Oral Oncol.*, **37**, 620–631.

Miyazawa, J., Mitoro, A., Kawashiri, S., Chada, K. K. and Imai, K. (2004) Expression of mesenchyme-specific gene HMGA2 in squamous cell carcinomas of the oral cavity. *Cancer Res.*, **64**, 2024–2029.

Mohan, K., Pearl, J. and Tian, J. (2013) Graphical models for inference with missing data. In *Advances in Neural Information Processing Systems 26* (eds C. J. C. Burges, M. Bottou, M. Welling, Z. Ghahramani and K. Q. Weinberger), pp. 1277–1285. Red Hook: Curran Associates.

Moll, U. M. and Petrenko, O. (2003) The MDM2-p53 interaction. *Molec. Cancer Res.*, **1**, 1001–1008.

Narita, M., Narita, M., Krizhanovsky, V., Nuñez, S., Chicas, A., Hearn, S. A., Myers, M. P. and Lowe, S. W. (2006) A novel role for high-mobility group a proteins in cellular senescence and heterochromatin formation. *Cell*, **126**, 503–514.

Ogata, H. and Goto, S. (2000) KEGG: Kyoto encyclopedia of genes and genomes. *Nucleic Acids Res.*, **28**, 27–30.

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

Parameswaran, J. and Burtness, B. (2018) P53 in head and neck squamous cell carcinoma. In *Molecular Determinants of Head and Neck Cancer* (eds B. Burtness and E. A. Golemis), ch. 9, pp. 249–274. Berlin: Springer.

Pearl, J. (2009) *Causality: Models, Reasoning, and Inference*, 2nd edn. Cambridge: Cambridge University Press.

Peltonen, J. K., Helppi, H. M., Paakko, P., Turpeenniemi-Hujanen, T. and Vahakangas, K. H. (2010) p53 in head and neck cancer: functional consequences and environmental implications of TP53 mutations. *Head Neck Oncol.*, **2**, article 36.

Pigeot, I., Sobotka, F., Kreiner, S. and Foraita, R. (2015) The uncertainty of a selected graphical model. *J. Appl. Statist.*, **42**, 2335–2352.

Piscuoglio, S., Zlobec, I., Pallante, P., Sepe, R., Esposito, F., Zimmermann, A., Diamantis, I., Terracciano, L., Fusco, A. and Karamitopoulou, E. (2012) HMGA1 and HMGA2 protein expression correlates with advanced tumour grade and lymph node metastasis in pancreatic adenocarcinoma. *Histopathology*, **60**, 397–404.

Richardson, T. and Spirtes, P. (2002) Ancestral graph Markov models. *Ann. Statist.*, **30**, 962–1030.

Robins, J. M., Scheines, R., Spirtes, P. and Wasserman, L. (2003) Uniform consistency in causal inference. *Biometrika*, **90**, 491–515.

de Roda Husman, A. M., Snijders, P. J., Stel, H. V., den Brule, A. J., Meijer, C. J. and Walboomers, J. M. (1995) Processing of long-stored archival cervical smears for human papillomavirus detection by the polymerase chain reaction. *Br. J. Cancer*, **72**, 412–417.

Rubin, D. B. (2004) *Multiple Imputation for Nonresponse in Surveys*. New York: Wiley.

Sachs, K., Perez, O., Pe'er, D., Lauffenburger, D. A. and Nolan, G. P. (2005) Causal protein-signaling networks derived from multiparameter single-cell data. *Science*, **308**, 523–529.

Scutari, M. (2010) Learning Bayesian networks with the bnlearn R package. *J. Statist. Softwr.*, **35**, 1–22.

Scutari, M., Vitolo, C. and Tucker, A. (2019) Learning Bayesian networks from big data with greedy search: computational complexity and efficient implementation. *Statist. Comput.*, **29**, 1095–1108.

Shi, X., Tian, B., Ma, W., Zhang, N., Qiao, Y., Li, X., Zhang, Y., Huang, B. and Lu, J. (2015) A novel anti-proliferative role of HMGA2 in induction of apoptosis through caspase 2 in primary human fibroblast cells. *Biosci. Rep.*, **35**, article e00169.

Spirtes, P., Glymour, C. and Scheines, R. (2000) *Causation, Prediction, and Search*, 2nd edn. Cambridge: MIT Press.

Spirtes, P. and Zhang, K. (2018) Search for causal models. In *Handbook of Graphical Models* (eds M. Maathuis, M. Drton, S. L. Lauritzen and M. Wainwright), 1st edn, ch. 18. Boca Raton: CRC Press.

Sterne, J. A. C., White, I. R., Carlin, J. B., Spratt, M., Royston, P., Kenward, M. G., Wood, A. M. and Carpenter, J. R. (2009) Multiple imputation for missing data in epidemiological and clinical research: potential and pitfalls. *Br. Med. J.*, **338**, article b2393.

Stewart, B. W. and Wild, C. P. (eds) (2014) *World Cancer Report 2014*. Lyon: World Health Organization and International Agency for Research on Cancer.

Strobl, E. V., Visweswaran, S. and Spirtes, P. L. (2018) Fast causal inference with non-random missingness by test-wise deletion. *Int. J. Data Sci. Analyt.*, **6**, 47–62.

Tsamardinos, I., Brown, L. E. and Aliferis, C. F. (2006) The max-min hill-climbing Bayesian network structure learning algorithm. *Mach. Learn.*, **65**, 31–78.

Tu, R., Zhang, C., Ackermann, P., Mohan, K., Glymour, C., Kjellström, H. and Zhang, K. (2019) Causal discovery in the presence of missing data. In *Proc. Mach. Learn. Res.* (eds K. Chaudhuri and M. Sugiyama), pp. 1762–1770.

Vogelstein, B., Lane, D. and Levine, A. J. (2000) Surfing the p53 network. *Nature*, **408**, 307–310.

Wei, J.-J., Wu, J., Luan, C., Yeldandi, A., Lee, P., Keh, P. and Liu, J. (2010) HMGA2: a potential biomarker complement to p53 for detection of early-stage high-grade papillary serous carcinoma in fallopian tubes. *Am. J. Surg. Pathol.*, **34**, 18–26.

Zhang, J. (2008a) Causal reasoning with ancestral graphs. *J. Mach. Learn. Res.*, **9**, 1437–1474.

Zhang, J. (2008b) On the completeness of orientation rules for causal discovery in the presence of latent confounders and selection bias. *Artif. Intell.*, **172**, 1873–1896.

Zhang, K., Schölkopf, B., Spirtes, P. and Glymour, C. (2018) Learning causality and causality-related learning: some recent progress. *Natn. Sci. Rev.*, **5**, 26–29.