SUPPLEMENTARY MATERIALS

Feasibility of lung cancer prediction from low-dose CT scan and smoking factors using causal models

Vineet K. Raghu[1,2], Wei Zhao[3], Jiantao Pu[3], Joseph K. Leader[3], Renwei Wang[4], James Herman[5], Jian-Min Yuan[4,6], Panayiotis V. Benos[1,2*], David O. Wilson[7]

[1]Department of Computer Science, University of Pittsburgh, Pittsburgh, PA

[2]Department of Computational and Systems Biology, University of Pittsburgh, Pittsburgh, PA

[3]Department of Radiology, University of Pittsburgh, Pittsburgh, PA

[4]UPMC Hillman Cancer Center, University of Pittsburgh, Pittsburgh, PA

[5]Division of Hematology, Oncology, Department of Medicine, University of Pittsburgh, Pittsburgh, PA

[6]Department of Epidemiology, Graduate School of Public Health, University of Pittsburgh, Pittsburgh, PA

[7]Division of Pulmonary, Allergy and Critical Care Medicine, School of Medicine, University of Pittsburgh, Pittsburgh, PA

Supported by National Institutes of Health (NIH) Grants U01HL137159 and R01LM012087 to PVB, R21CA197493 to JP, T32CA082084 to VKR, the University of Pittsburgh Cancer Institute's Specialized Program of Research Excellence (SPORE) in Lung Cancer (NCI P50CA90440) and the Cancer Center Core Grant (NCI 2P30 CA047904)

Key words: lung cancer risk, low dose CT, cancer screening

[*]Correspondence and requests for reprints should be addressed to:
Panayiotis V. Benos, PhD
Department of Computational and Systems Biology
Suite 3064, Biomedical Sciences Tower 3
3501 Fifth Avenue, Pittsburgh, PA 15260.
E-mail: benos@pitt.edu

SUPPLEMENTARY METHODS

Probabilistic mixed graphical models used in this paper

Probabilistic graphical models (PGMs) are a robust way to represent the dependencies and conditional dependencies in the data and they can also be used to build predictive models. Until recently, PGMs could learn graphs only if all data were the same type (continuous, discrete). We developed MGM-FCI-MAX (1), an extension of standard PGMs that allows for the analysis of datasets with mixed data types, and used it to learn the informative features for lung cancer identification from multi-scale data (demographics, CT scans, etc.). MGM-FCI-MAX works in two steps. First, it calculates the undirected graph, which is equivalent to partial correlation or graphical lasso for mixed data types. We do so by modeling the likelihood of all variables in a composite model (**Equation 1**).

$$p(x, y, \Theta) \propto exp\left(\sum_{\omega=1}^{p}\sum_{\varphi=1}^{p}\beta_{\omega\varphi}\, x_\omega x_\varphi + \sum_{\omega=1}^{p}\alpha_\omega x_\omega + \sum_{\omega=1}^{p}\sum_{\upsilon=1}^{q}\rho_{\omega\upsilon}(y_\upsilon)\, x_\omega + \sum_{\upsilon=1}^{q}\sum_{\zeta=1}^{q}\Upsilon_{\zeta\upsilon}(y_\upsilon, y_\zeta)\right). \quad (Eq.1)$$

In this equation, x refers to one of p continuous variables, and y refers to one of q categorical variables. $\beta_{\omega\varphi}$ is the linear coefficient between two continuous variables, $\rho_{\omega\upsilon}$is a vector of coefficients that represents the interaction between each category of y_υ and x_ω, and $\Upsilon_{\zeta\upsilon}$ is a matrix of coefficients between pairs of categorical variables: y_υ, y_ζ. To ensure a sparse graph and avoid overfitting we use separate regularization parameters for each type of edge (λ_{CC}, λ_{DD}, λ_{CD} for edges between two continuous, two discrete or a continuous and discrete variables) (**Equation 2**):

$$argmin_\Theta\, \tilde{l}(\Theta) + \lambda_{cc}\sum_{\varphi<\omega}\left|\beta_{\omega\varphi}\right|_1 + \lambda_{cd}\sum_{\omega,\upsilon}\|\rho_{\omega\upsilon}\|_2 + \lambda_{dd}\sum_{\zeta<\upsilon}\left\|\Upsilon_{\zeta\upsilon}\right\|_F \quad (Eq.2)$$

Here, $\tilde{l}(\Theta)$ refers to the negative log-likelihood of the model, which we minimize using a proximal gradient approach as the original authors did. In addition, to optimize the λ parameters of the model, we use the StEPS procedure proposed before (2).

The second step of MGM-FCI-MAX consists of orienting edges considering that unmeasured confounders (latent variables) might influence the variables in the dataset. This is an important improvement, especially for analysis of clinical datasets, because most of them are expected to have many unmeasured relevant variables due to technical inability to measure them or lack of knowledge of their importance in this disease. MGM-FCI-MAX is more accurate than other methods, and it has demonstrated usefulness in biomedical data (3).

The output of the algorithm is a graphical causal model where edges have three possible endpoints. An edge of the form ("A→B") suggests that B is not a cause of A (">" means not a cause), and A is a cause of B ("-" means cause). An edge ("A<-->B") suggests that neither A nor B is a cause of the other, that is, a latent variable causes both. Finally, an edge of the form ("Ao-oB") suggests that both endpoints are inconclusive from the data. We note that in high dimensional datasets (small sample size, large number of variables) all these algorithms are not as accurate in inferring the causal orientation as they are in inferring the presence of an edge(4).

REFERENCES

1. Raghu VK, Ramsey JD, Morris A, Manatakis DV, Sprites P, Chrysanthis PK*, et al.* Comparison of strategies for scalable causal discovery of latent variable models from mixed data. International Journal of Data Science and Analytics **2018**:1-13 doi 10.1007/s41060-018-0104-3.

2. Sedgewick AJ, Shi I, Donovan RM, Benos PV. Learning mixed graphical models with separate sparsity parameters and stability-based model selection. BMC Bioinformatics **2016**;17 Suppl 5:175 doi 10.1186/s12859-016-1039-0.
3. Raghu VK, Ramsey JD, Morris A, Manatakis DV, Spirtes P, Chrysanthis PK*, et al.* Comparison of Strategies for Scalable Causal Discovery of Latent Variable Models from Mixed Data. 2017 ACM SIGKDD. Halifax, Nova Scotia, Canada: Springer; 2017.
4. Raghu VK, Ramsey JD, Morris A, Manatakis DV, Sprites P, Chrysanthis PK*, et al.* Comparison of strategies for scalable causal discovery of latent variable models from mixed data. Int J Data Sci Anal **2018**;6(1):33-45 doi 10.1007/s41060-018-0104-3.

SUPPLEMENTARY TABLE S1. The PLuSS cohort variables (training cohort) used in the analysis and the development of LCCM. Bin: binary; Cat: categorical; Cont: continuous; Num: numerical;

	Type	Definition
Diagnosis		
Cancer status	Bin	0: Benign nodule, 1: Cancer
Demographics		
Age	Num	Subject's age
BMI	Cont	Body Mass Index
Education	Cat	Five categories describing the highest educational level
Sex	Bin	Male, Female
Comorbidites		
Bronchitis	Bin	0: No, 1: Yes
Emphysema	Bin	0: No, 1: Yes
Pack Years	Num	Average number of cigarettes smoked daily divided by 20 X number of years smoked
Years since quit smoking	Num	Years since subject quit smoking
CT scan		
Area (cm^2)	Cont	the surface area of a nodule
Cavity ratio	Cont	the ratio between the cavity volume and the nodule volume
Ground glass opacity	Cont	the difference between the nodule density and the air density
Irregularity	Cont	the ratio between the surface area and the volume of a nodule
Max diameter (cm)	Cont	the largest distance of two points on the nodule surface
Mean diameter (cm)	Cont	the mean distance of two points on the nodule surface
Mean diameter solid portion (cc)	Cont	the mean diameter of the solid part of a nodule
Mean intensity (HU)	Cont	the mean intensity of a nodule in Hounsfield Units
Mean vessel intensity (HU)	Cont	the mean intensity of the surrounding vessels in Hounsfield Units
Nodule type	Bin	solid or non-solid
Nodules, Number of	Num	number of nodules detected in this subject
Vessel volume (mL)	Cont	the volume of the vessels surrounding a nodule
Vessel, Number of	Num	Number of vessels around the examined nodule
Volume (mL)	Cont	nodule volume
Volume cal score (mm^3)	Cont	nodule calcification volume

SUPPLEMENTARY TABLE S2. Characteristics of all the nodules in the validation cohort.

Features	Lung cancer (n = 39)	Benign nodules (n = 87)	P value†
Male, n (%)	22 (56)	52 (60)	0.874
Age, mean, years (SD)	65.28 (9.14)	67.95 (8.42)	0.125
Current smoker, n (%)	32 (82)	41 (47)	<0.001
Pack-Years, mean (SD)*	50.45 (23.25)	58.48 (27.04)	0.099
Years since quit smoking, mean (SD)	0.538 (1.59)	2.76 (4.17)	<0.001
Nodule size in diameter (mm), mean (SD)	18.69 (6.24)	10.81 (4.51)	<0.001
Nodule number, n (%) °			0.152
Solid	25 (78)	52 (62)	
Non-solid/mixed	7 (22)	32 (38)	
Vessel number, mean (SD)	15.92 (11.85)	3.59 (3.80)	<0.001

Abbreviations: SD, standard deviation

† Two-sided *p*-values were based on *t* test and chi-square test for continuous and categorical variables, respectively.

° Nodule type was unmeasured for nine subjects (seven with cancer)

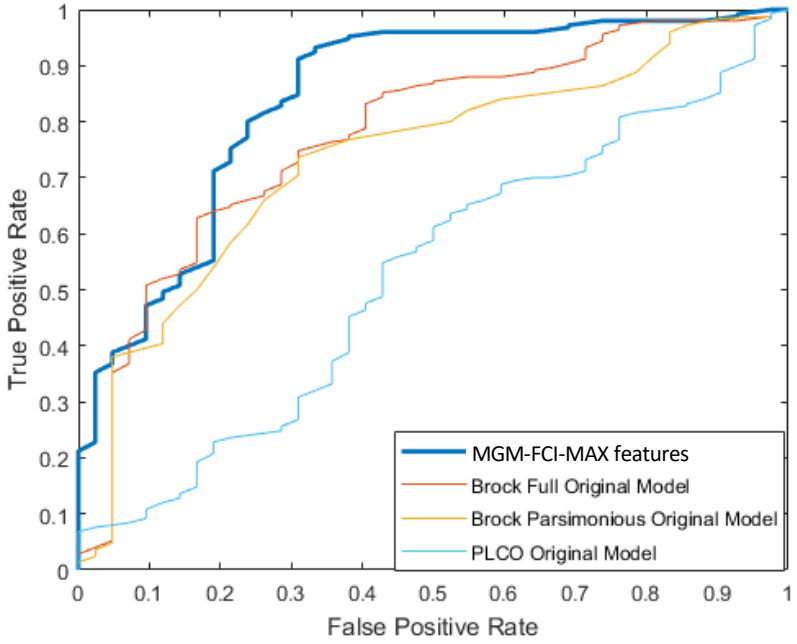
* Pack-Years was unmeasured for three subjects (two with cancer)

SUPPLEMENTARY TABLE S3. The features included in the MGM-FCI-MAX Markov blanket around "cancer status" in each of the 10X cross-validation rounds.

1	✓	✓	✓		
2	✓	✓	✓		
3	✓	✓	✓		
4	✓	✓	✓	✓	
5	✓	✓	✓		✓
6	✓	✓			
7	✓	✓			
8	✓	✓	✓		
9	✓	✓	✓		
10	✓	✓	✓		
Total	10	10	8	1	1

SUPPLEMENTARY FIGURE S1. Comparison of MGM-FCI-MAX derived to published lung cancer prediction models with their original coefficients on the training cohort. **(A)** Receiver operating characteristic (ROC) curves from the cross-validation results on the training cohort for LCCM and various published models with *their original coefficients*. ROC curves were computed using nested 10-fold cross validation, and model discrimination was measured by area under the ROC curve (AUC). **(B)** Detailed numerical results of model comparison. p-values are computed via a paired t-test between our *Lung Cancer Causal Model* (LCCM) and previously published models.

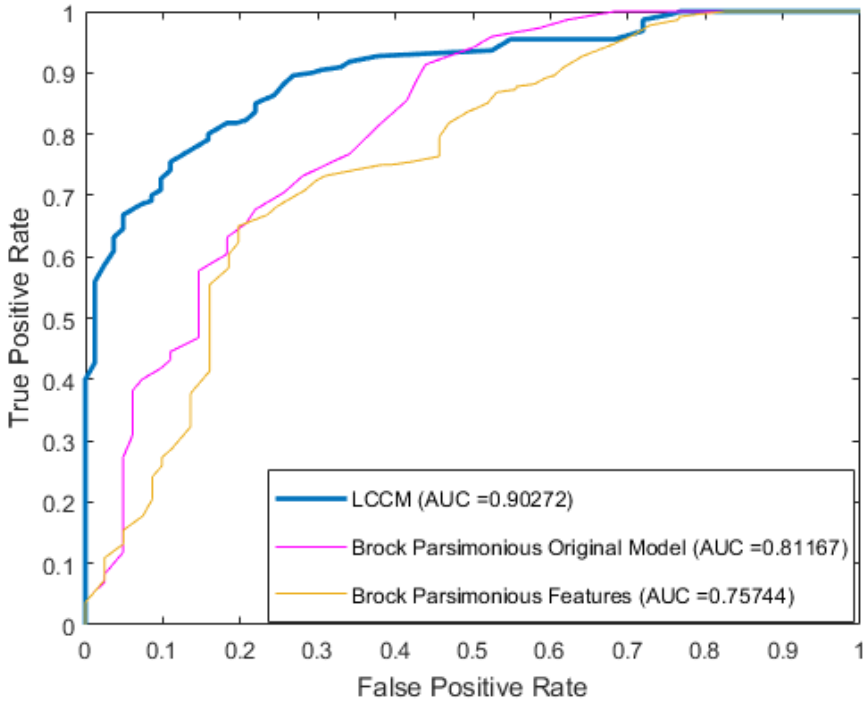
A



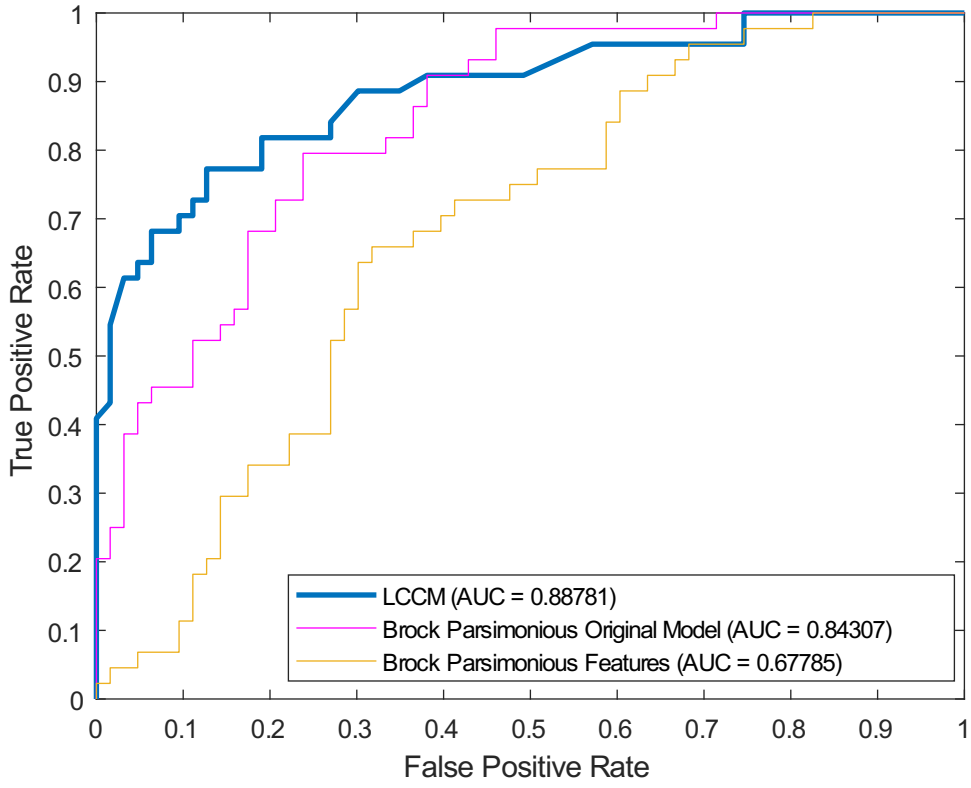
B

Model	No. of Features	AUC [25%, 75%]	p-value	Features Used
MGM-FCI-MAX features	3	0.882 (0.786, 1.000)	-	**Smoking:** Years Quit **Radiographic:** Nodule Count, Vessel Number
Brock Full (Original)	8	0.768 (0.650,0.917)	0.11	**Demographics:** Age, Sex, Family History Ca **Comorbidities:** Emphysema **Radiographic:** Nodule Size, Nodule Type, Nodule Location, Nodule Count
Brock Parsimonious (Original)	3	0.712 (0.500,0.857)	0.05	**Demographics**: Sex **Radiographic:** Nodule Location, Nodule Size
PLCO (Original)	10	0.466 (0.214,0.640)	<0.01	**Demographics**: BMI, Education, Family History Ca, Race **Comorbidities**: Ca History, COPD **Smoking**: Duration, Intensity, Smoking Status, Years Quit

SUPPLEMENTARY FIGURE S2. Receiver operating characteristic (ROC) curves from the external (validation) cohort results for LCCM and two Brock Parsimonious models: the one with the parameters in the original publication ("Brock Parsimonious Original Model") and the re-trained one with coefficients estimated in the same training cohort as LCCM (Brock Parsimonious Features"). The *p*-values of the difference in AUC are significant for both Brock models (*p*-value <0.01 and 0.0176 for the re-trained and original models, respectively).



SUPPLEMENTARY FIGURE S3. Receiver operating characteristic (ROC) curves from the external (validation) cohort using all benign nodules. The predicted probabilities for each model correspond to the probability that a subject has cancer, based upon the highest predicted probability for all of the nodules for this subject. Results are shown for LCCM and two Brock Parsimonious models: the one with the parameters in the original publication ("Brock Parsimonious Original Model") and the re-trained one with coefficients estimated in the same training cohort as LCCM (Brock Parsimonious Features"). The *p*-values of the difference in AUC is significant for the retrained Brock model (p <0.01) and not significant for the original Brock model (p = 0.22245)



SUPPLEMENTARY FIGURE S4. Receiver operating characteristic (ROC) curves from the external (validation) cohort results for all benign nodules <3 cm. Comparison between LCCM and two Brock Parsimonious models: the one with the parameters in the original publication ("Brock Parsimonious Original Model") and the re-trained one with coefficients estimated in the same training cohort as LCCM (Brock Parsimonious Features"). The *p*-value of the difference in AUC is significant for the retrained Brock model ($p<0.01$), but not the Brock Original Model ($p=0.263$).