Treatment of Missing Data in Bayesian Structural Learning: A Simulation Study for Social Science: a case-study of Antimicrobial resistance

Xueija Ke, Madeleine Clarkson, Katherine Keenan & V Anne Smith

University of St Andrews

Scottish Graduate School of Social Science

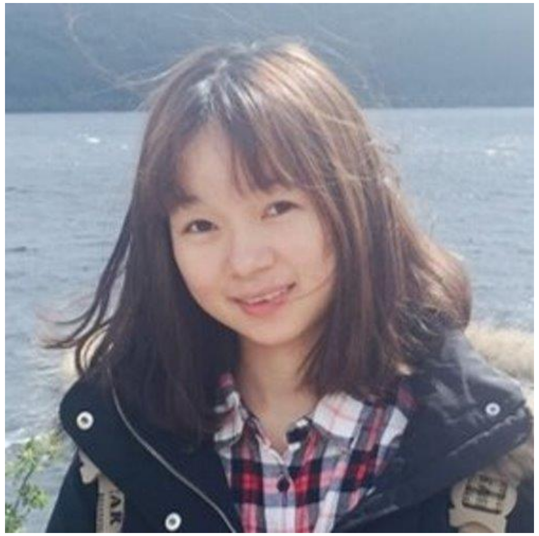
UK Research and Innovation

Holistic Approach To Unravel Antibacterial Resistance in East Africa

Outline

- Demonstrate how Bayesian belief networks(BBNs) can be used and interpretate for social science data
- Introduce antimicrobial drug resistance(AMR) as a bio-socially complex phenomenon
- Demonstrate how Bayesian logic is useful for understanding the AMR phenomenon
- Summarise results from a literature review of BBNs in AMR and antibiotic use

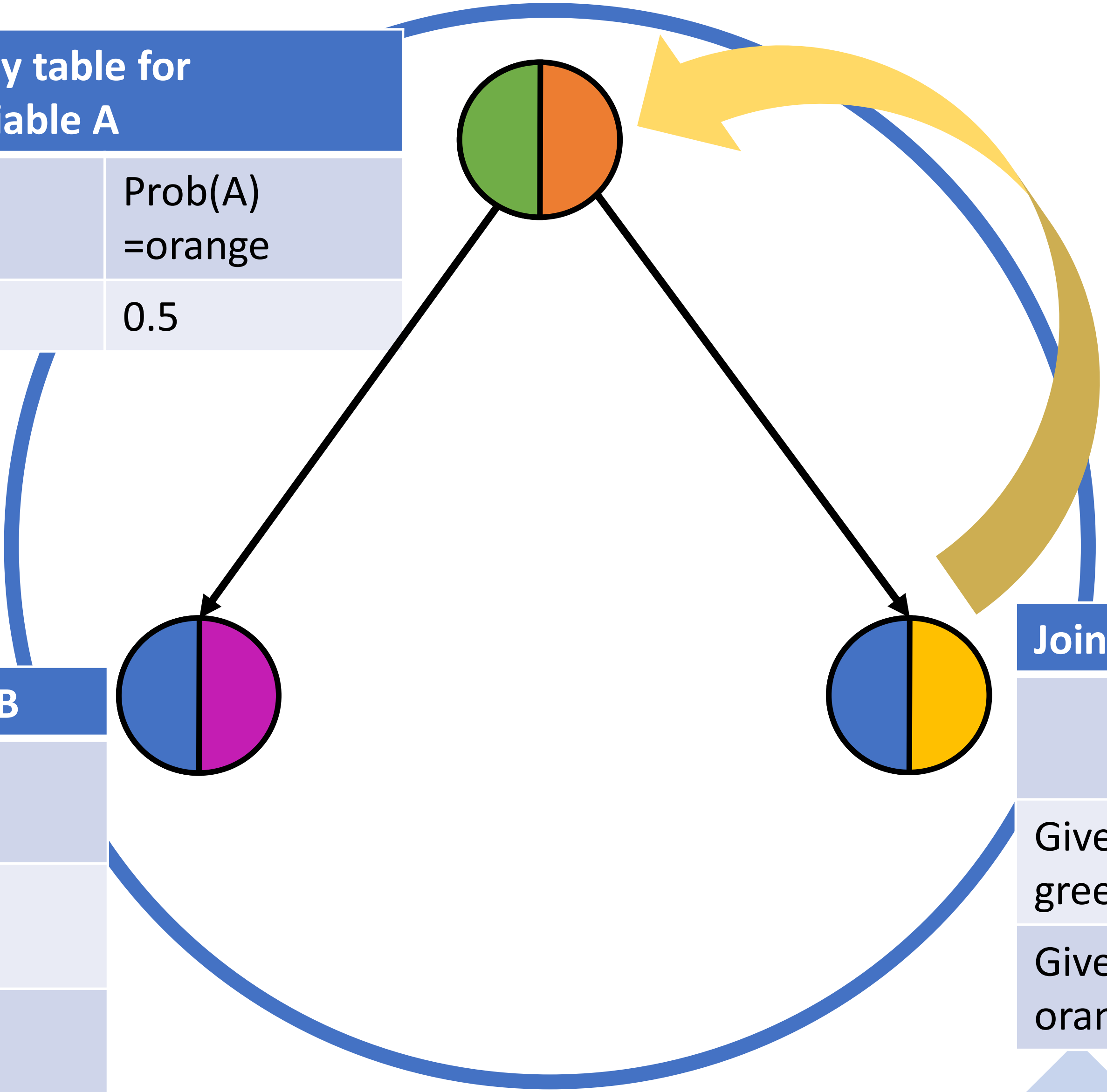
- Introduce missing data and three missing mechanisms
- A brief review on how BBNs can be used for dealing with missing data
- Demonstrate a simulation study on comparing the performance of two popular approaches for missing data
- Demonstrate an application on real data - case study of AMR



How can Bayesian Belief Networks be used in social science?

1. **Variable inference**
2. **Parameter inference**
3. **Structure-learning**

Variable inference



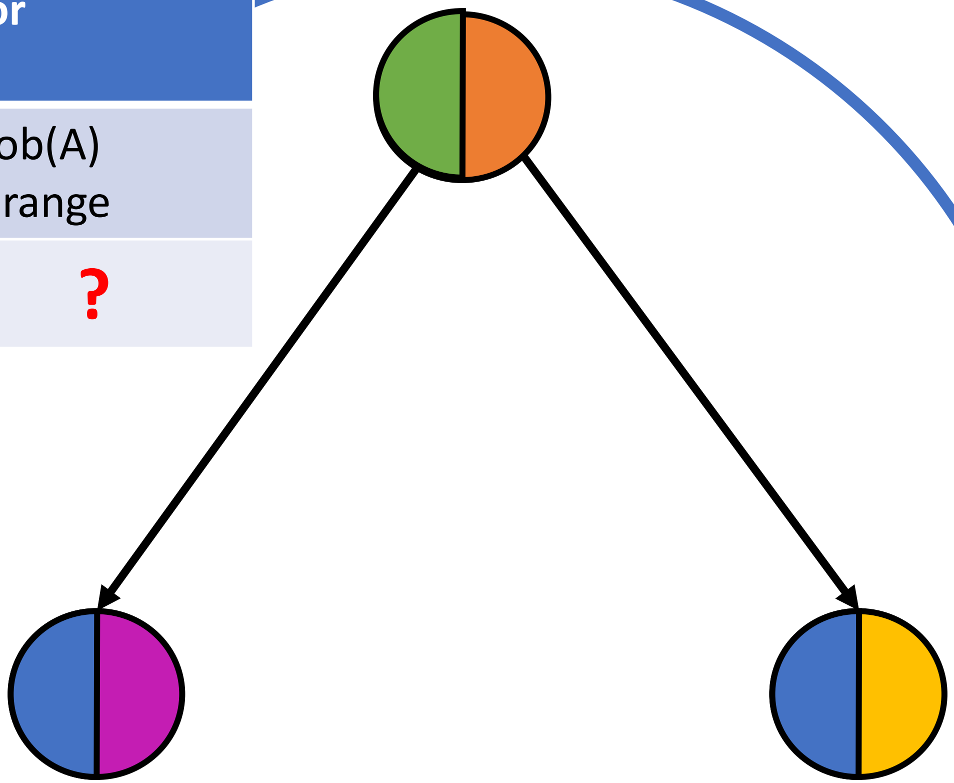
Probability table for node/variable A	
Prob (A) =green	Prob(A) =orange
0.5	0.5

Joint Probability table for mode B		
	Prob(B) =pink	Prob(B) =blue
Given A= green	0.6	0.4
Given A= orange	0.3	0.7

Joint Probability table for mode C		
	Prob(C) =yellow	Prob(C) =blue
Given A= green	0.1	0.9
Given A= orange	0.9	0.1

Parameter inference

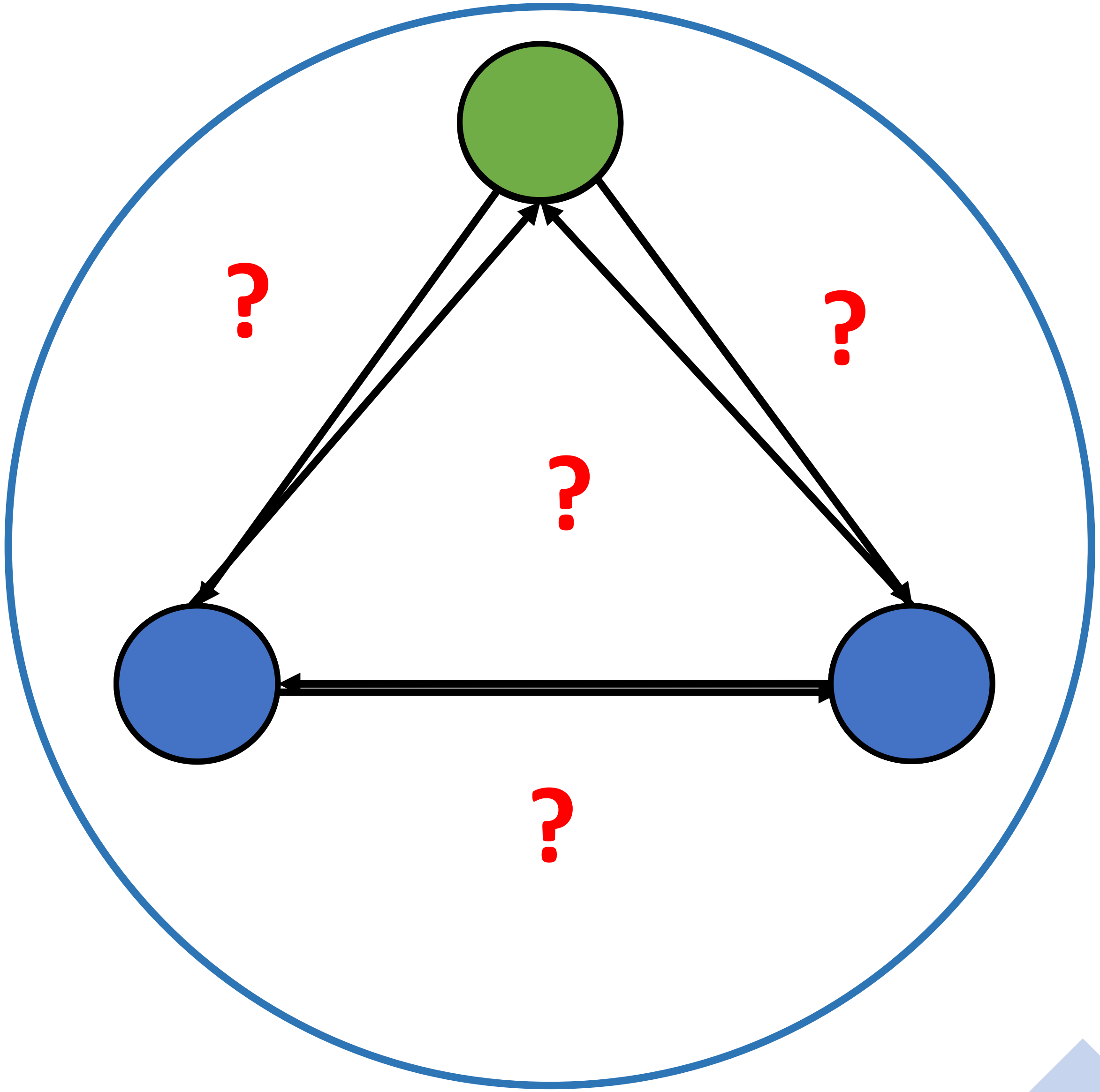
Probability table for node/variable A	
Prob (A) =green	Prob(A) =orange
?	?



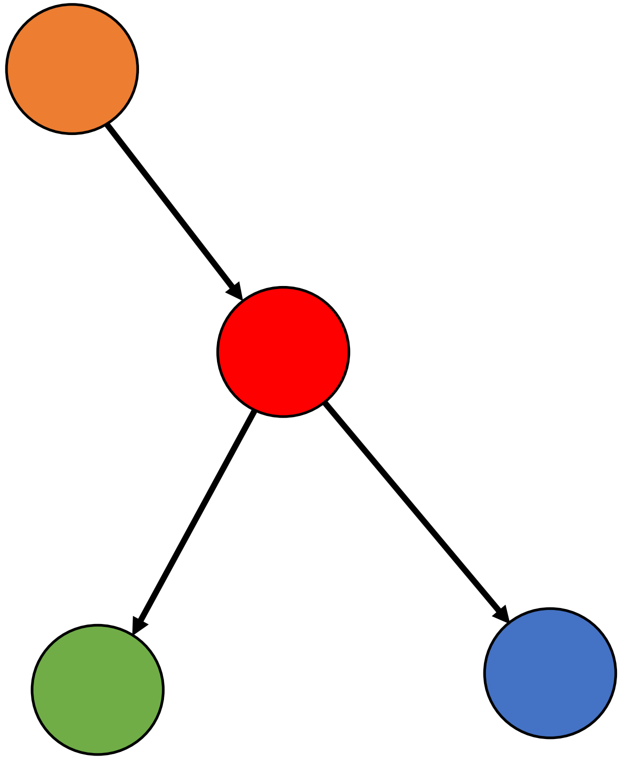
Joint Probability table for mode B		
	Prob(B) =pink	Prob(B) =blue
Given A= green	?	?
Given A= orange	?	?

Joint Probability table for mode C		
	Prob(C) =yellow	Prob(C) =blue
Given A= green	?	?
Given A= orange	?	?

Structure-learning

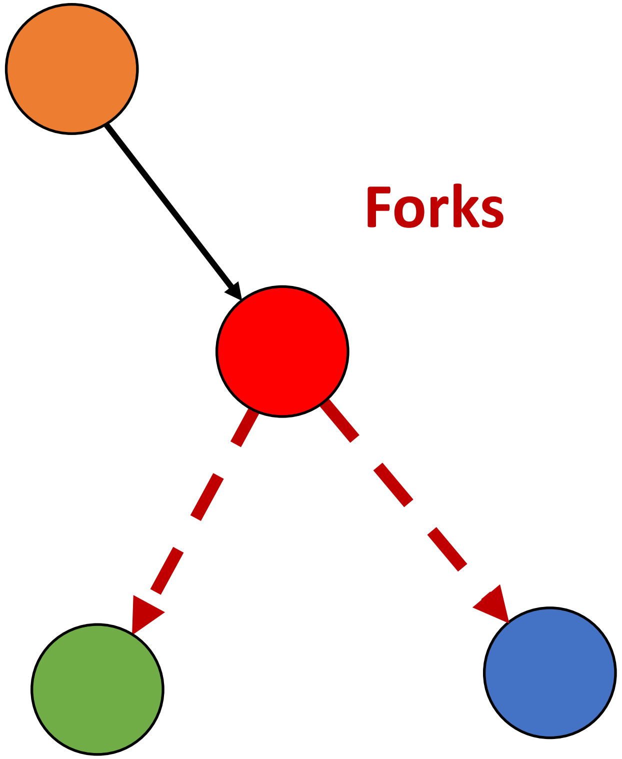


Bayesian Belief networks provide "actionable motifs"[1] which guide social science inference, interpretation and further investigation



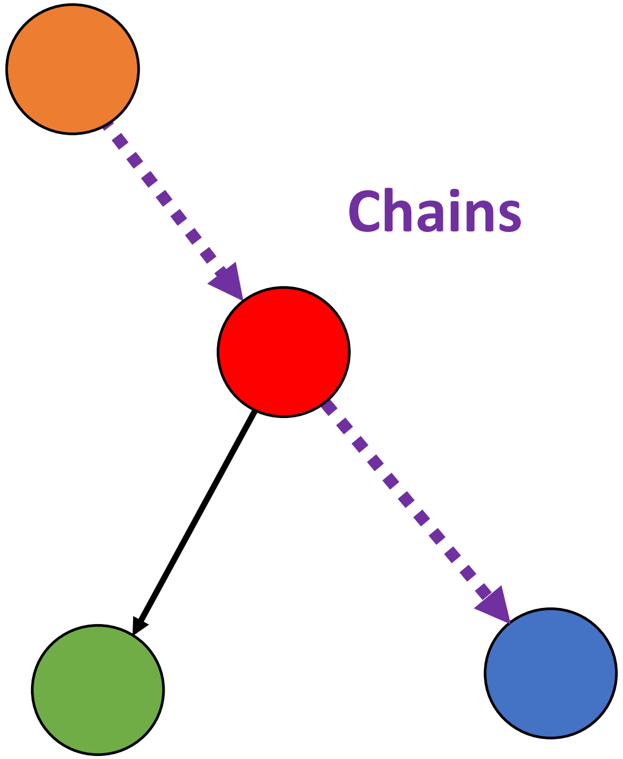
[1] Sethi, T. *et al.* (2018) 'Stewarding antibiotic stewardship in intensive care units with Bayesian artificial intelligence [version 1; peer review: 2 approved with reservations]', *Wellcome Open Research*, 3. doi: 10.12688/wellcomeopenres.14629.1.

Bayesian Belief networks provide "actionable motifs"[1] which guide social science inference, interpretation and further investigation



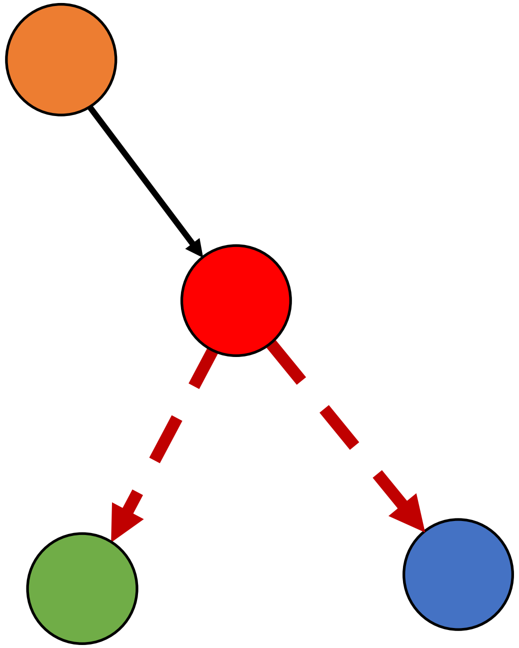
[1] Sethi, T. *et al.* (2018) 'Stewarding antibiotic stewardship in intensive care units with Bayesian artificial intelligence [version 1; peer review: 2 approved with reservations]', *Wellcome Open Research*, 3. doi: 10.12688/wellcomeopenres.14629.1.

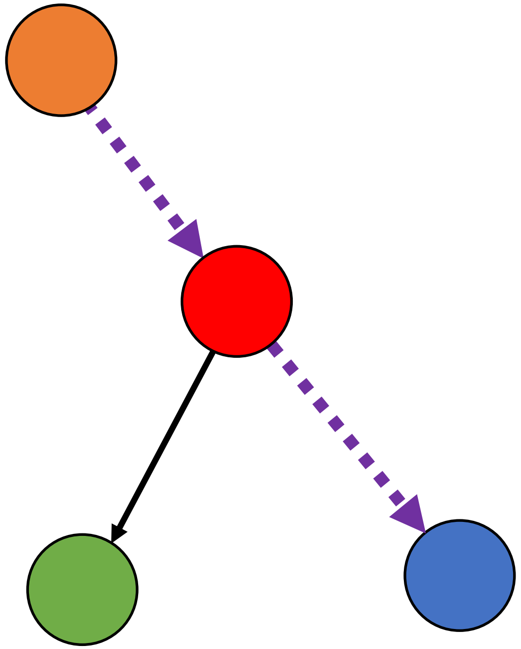
Bayesian Belief networks provide "actionable motifs"[1] which guide social science inference, interpretation and further investigation



[1] Sethi, T. *et al.* (2018) ‘Stewarding antibiotic stewardship in intensive care units with Bayesian artificial intelligence [version 1; peer review: 2 approved with reservations]’, *Wellcome Open Research*, 3. doi: 10.12688/wellcomeopenres.14629.1.

Confounders and mediators





Confounder variable
Is a variable which influences two variables causing a spurious association to between them[2], within BBNs is represented by ***forks*** in a directed graphical network[1]

Mediator variable
Is a variable which explains the process through which two variables are related[3], within BBNs are represented by nodes within a *chain of arcs* in a directed graphical network1

1. Sethi, T. *et al.* (2018) 'Stewarding antibiotic stewardship in intensive care units with Bayesian artificial intelligence [version 1; peer review: 2 approved with reservations]', *Wellcome Open Research*, 3. doi: 10.12688/wellcomeopenres.14629.1.

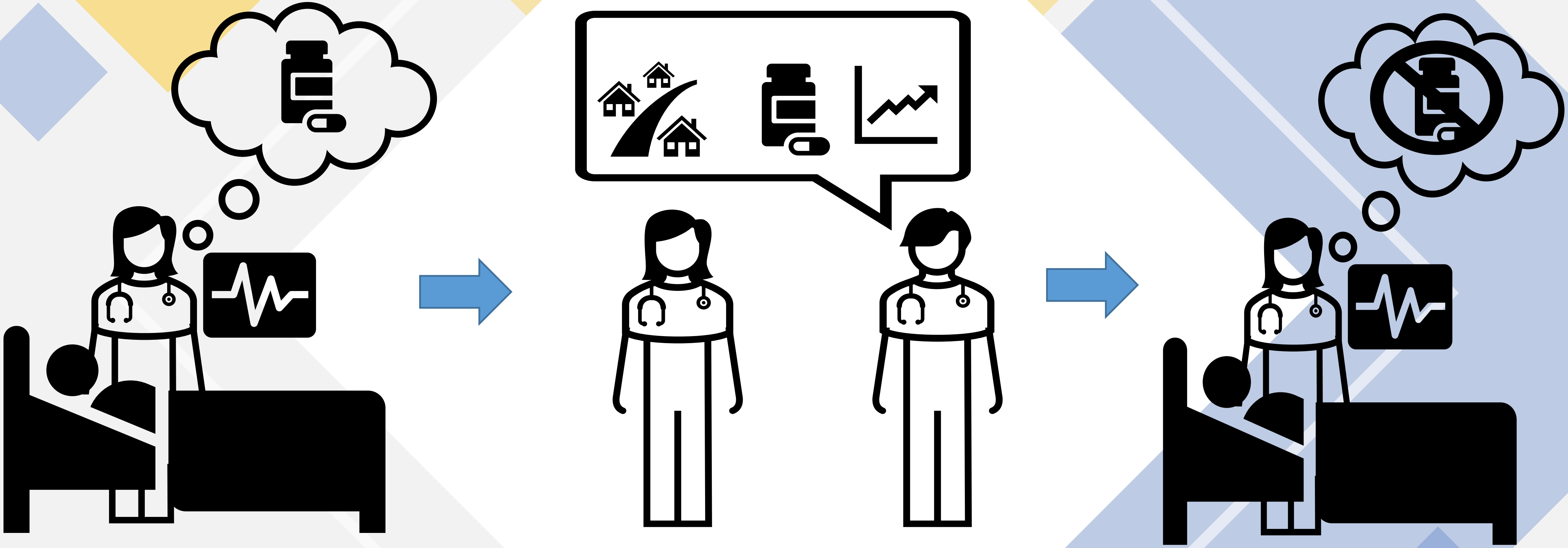
2. Pearl, J., (2009). [Simpson's Paradox](), Confounding, and Collapsibility In *Causality: Models, Reasoning and Inference* (2nd ed.). New York : Cambridge University Press.

3. Pritha Bhandari,2021 Mediator *vs Moderator variables* [accessed online] https://www.scribbr.com/methodology/mediator-vs-moderator/

Antimicrobial resistance

- **Antimicrobial resistance → Evade or survive** treatment
- Resistance is a complex issue:
  - Exposure
  - vertical & horizontal gene transfer [Vikesland et al,2020]
  - Animals - Environment - Human
- **Biosocially complex**
- BBNs → complexity
- limited Use in the AMR literature

Bayesian logic applied to AMR



$$P(A|B) = \frac{P(B|A)P(A)}{P(B|A)P(A) + P(B|\text{not } A)P(\text{not } A)}$$

Review of the use of BBN applications in the AMR and antibiotic use literature

Iterative scoping review[4]:

- Literature landscape - terms "Bayes", "AMR" and "antibiotics", "antimicrobial resistance"
- How and What
- ~~Bayesian statistical applications~~
- Boolean searches – *Pearl growing*
- citation tracking

[4] Martin, G. P. *et al.* (2020) 'Toward a framework for the design, implementation, and reporting of methodology scoping reviews', *Journal of Clinical Epidemiology*. Elsevier Inc, 127, pp. 191–197. doi: 10.1016/j.jclinepi.2020.07.014.

Review of BBNs in the field of AMR

Paper	Main purpose
(Ge *et al.*, 2014)	Analyse the association between socioeconomic causal factors of antibiotic use in livestock.
(Ludwig e	Analyse the associations resistance patterns in pig farming (*E.coli)*
(Hartnack	Analyse the associations in chicken farming (*Salmonella spp.)*
(Hidalgo *et al.*, 2015)	Analyse the associations in chicken retail (*E. faecalis*)
(Cherry *et al.*, 2021)	Analyse the associations of cross-resistance patterns and antibiotic use in UTI patients
(Sethi	Analyse the (Paediatric ICU) antibiotic sensitivities to develop a tool to replace antibiograms
(Wu *et al.*, 2020)	Develops a tool for clinicians to appropriately prescribe antibiotics & predict causative pathogen (osteomyelitis)
TREAT CPN	Develops a tool for clinicians to appropriately prescribe antibiotics (multiple pathogens)
(Lucas *et a*	Develops a tool for clinicians to appropriately prescribe antibiotics (pneumonia in the ICU)
(Leibovici	Develops a tool for clinicians to appropriately prescribe antibiotics (UTI patients)
(Beuscart *et al.*, 1999)	Develops a tool for clinicians to appropriately prescribe antibiotics (UTI patients)
(Andreassen *et al.* 1999)	Decision tool to balance therapeutic benefit and cost of antibiotics (UTI patients)

analysis
n= 5

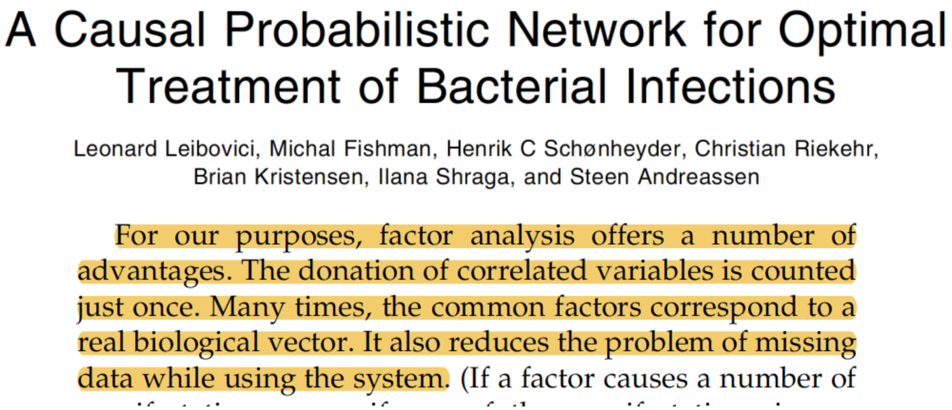
meat production

Both n= 1

Hospital setting

Decision
tool n= 6

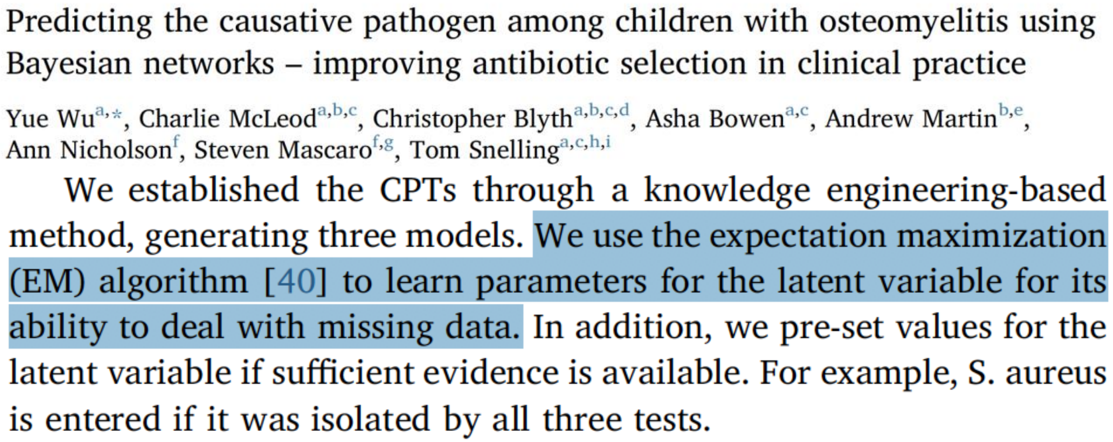
Review of BBNs in the field of AMR: dealing with incomplete data with non-learnt structures



A Causal Probabilistic Network for Optimal Treatment of Bacterial Infections

Leonard Leibovici, Michal Fishman, Henrik C Schønheyder, Christian Riekehr, Brian Kristensen, Ilana Shraga, and Steen Andreassen

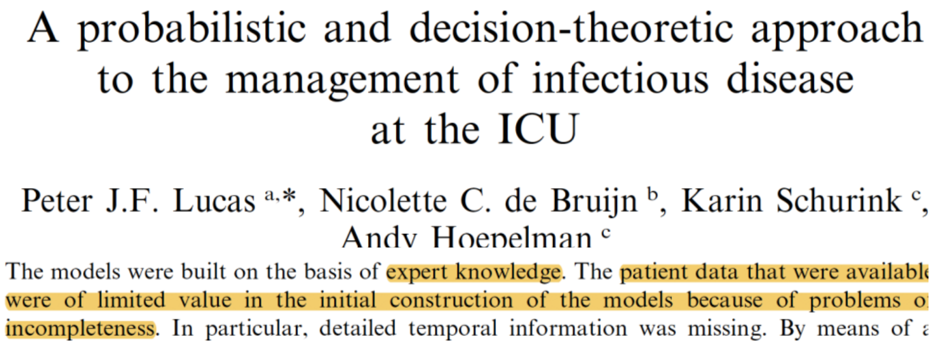
For our purposes, factor analysis offers a number of advantages. The donation of correlated variables is counted just once. Many times, the common factors correspond to a real biological vector. It also reduces the problem of missing data while using the system. (If a factor causes a number of



Predicting the causative pathogen among children with osteomyelitis using Bayesian networks – improving antibiotic selection in clinical practice

Yue Wu[a,*], Charlie McLeod[a,b,c], Christopher Blyth[a,b,c,d], Asha Bowen[a,c], Andrew Martin[b,e], Ann Nicholson[f], Steven Mascaro[f,g], Tom Snelling[a,c,h,i]

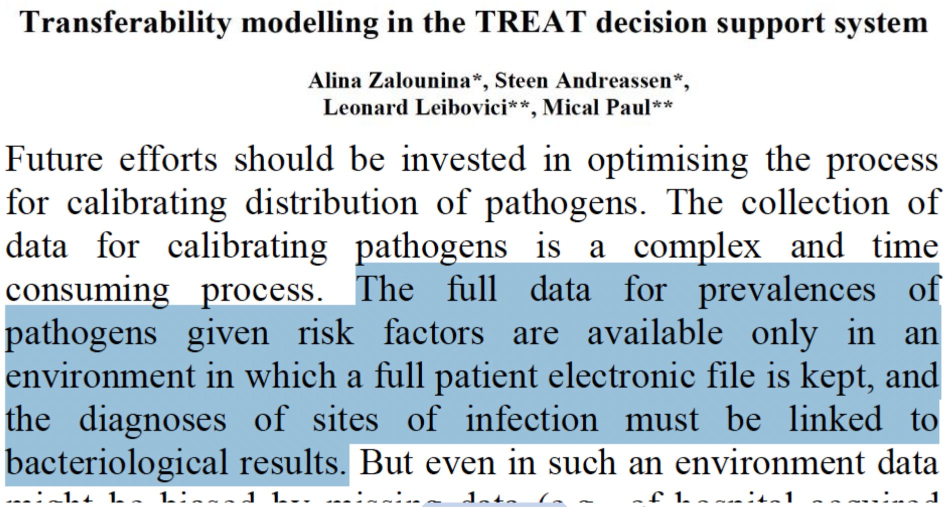
We established the CPTs through a knowledge engineering-based method, generating three models. We use the expectation maximization (EM) algorithm [40] to learn parameters for the latent variable for its ability to deal with missing data. In addition, we pre-set values for the latent variable if sufficient evidence is available. For example, S. aureus is entered if it was isolated by all three tests.



A probabilistic and decision-theoretic approach to the management of infectious disease at the ICU

Peter J.F. Lucas[a,*], Nicolette C. de Bruijn[b], Karin Schurink[c], Andy Hoepelman[c]

The models were built on the basis of expert knowledge. The patient data that were availabl were of limited value in the initial construction of the models because of problems o incompleteness. In particular, detailed temporal information was missing. By means of a

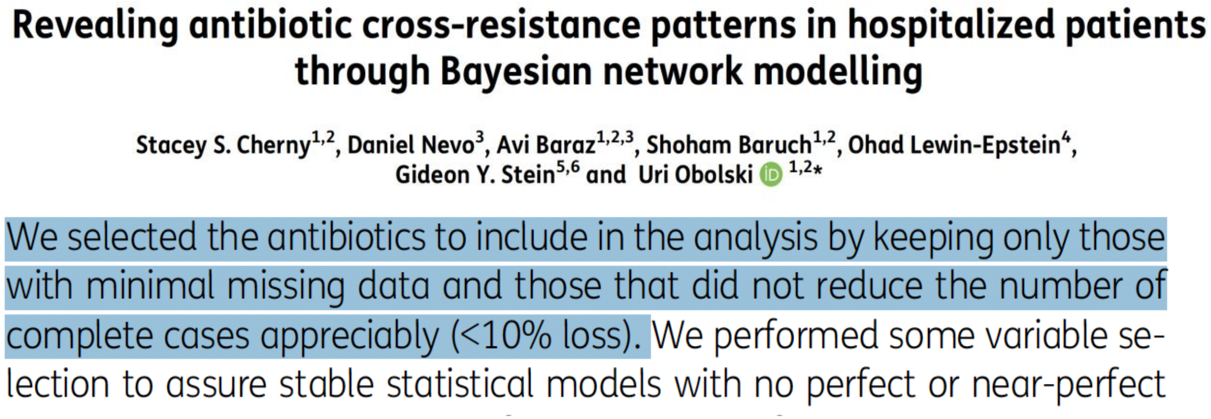


Transferability modelling in the TREAT decision support system

Alina Zalounina*, Steen Andreassen*, Leonard Leibovici**, Mical Paul**

Future efforts should be invested in optimising the process for calibrating distribution of pathogens. The collection of data for calibrating pathogens is a complex and time consuming process. The full data for prevalences of pathogens given risk factors are available only in an environment in which a full patient electronic file is kept, and the diagnoses of sites of infection must be linked to bacteriological results. But even in such an environment data might be biased by missing data (e.g. of hospital acquired

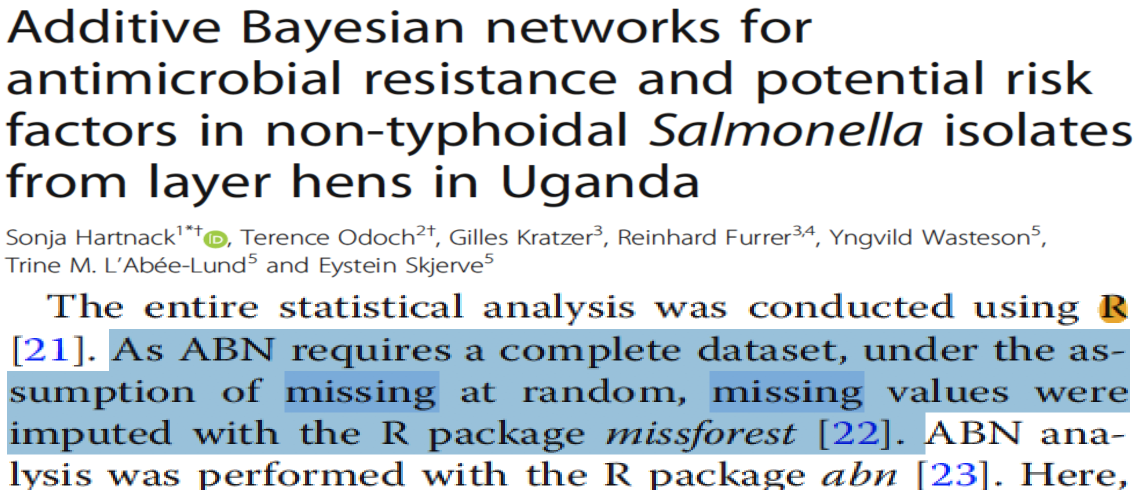
Review of BBNs in the field of AMR: dealing with incomplete data with learnt structures



Revealing antibiotic cross-resistance patterns in hospitalized patients through Bayesian network modelling

Stacey S. Cherny[1,2], Daniel Nevo[3], Avi Baraz[1,2,3], Shoham Baruch[1,2], Ohad Lewin-Epstein[4], Gideon Y. Stein[5,6] and Uri Obolski [1,2*]

We selected the antibiotics to include in the analysis by keeping only those with minimal missing data and those that did not reduce the number of complete cases appreciably (<10% loss). We performed some variable selection to assure stable statistical models with no perfect or near-perfect

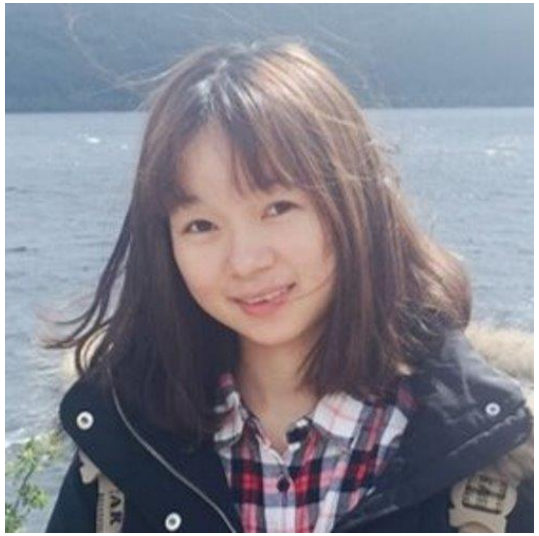


Additive Bayesian networks for antimicrobial resistance and potential risk factors in non-typhoidal *Salmonella* isolates from layer hens in Uganda

Sonja Hartnack[1*†], Terence Odoch[2†], Gilles Kratzer[3], Reinhard Furrer[3,4], Yngvild Wasteson[5], Trine M. L'Abée-Lund[5] and Eystein Skjerve[5]

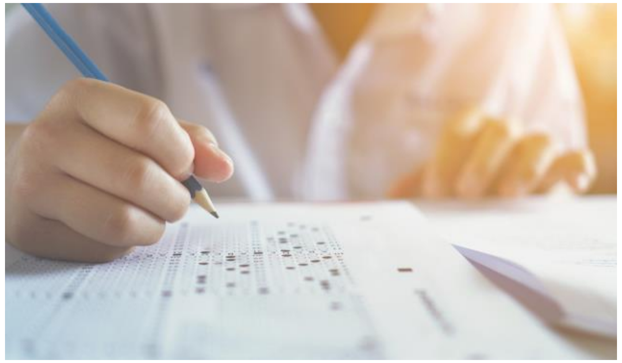
The entire statistical analysis was conducted using R [21]. As ABN requires a complete dataset, under the assumption of missing at random, missing values were imputed with the R package *missforest* [22]. ABN analysis was performed with the R package *abn* [23]. Here,

- Introduce missing data and three missing mechanisms
- A brief review on how BBNs can be used for dealing with missing data
- Demonstrate a simulation study on comparing the performance of two popular approaches for missing data
- Demonstrate an application on real data - case study of AMR



Missing Data

A	B	C
5	10	?
4	?	6



Missing Mechanisms

- MCAR - Missing Completely at Random (rare)

-- missingness is **unrelated** to unobserved & observed responses

- MAR – Missing at Random (common)

-- missingness is **unrelated** to unobserved response but **related** to observed response

- MNAR – Missing Not at Random (difficult to detect)

-- missingness is **related to both** unobserved and observed responses

Bayesian Networks & Missing Data

- Structure learning from incomplete data

-- data completion & refinement + standard learning algorithms & scores (e.g., Structural EM algorithm)

-- approximate BIC scores & marginal likelihood P(D|G) (e.g., variational-Bayesian EM algorithm)

- Parameter learning from incomplete data given a known structure (assume MCAR or MAR)

-- data augmentation (DA; Tanner &Wong, 1987)

-- expectation–maximisation algorithm (EM; Lauritzen, 1995)

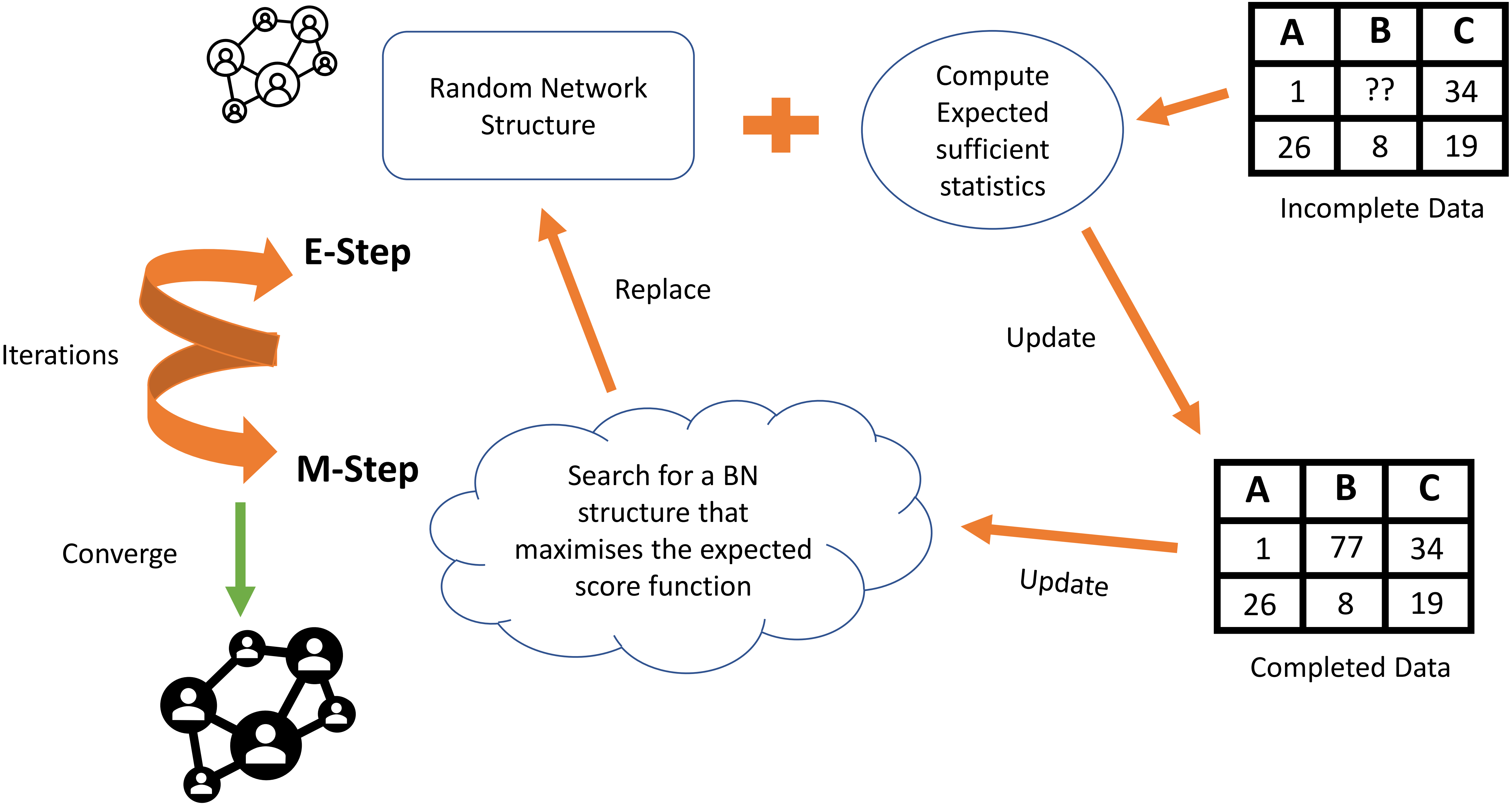
-- Bound and Collapse (also robust for MNAR data) [BC; Ramoni & Sebastiani, 1997]

-- robust Bayesian estimator (RBE; Ramoni & Sebastiani)

-- simple imputation methods (Oni´sko, Druzdzel, & Wasyluk, 2002)

Ref: Bayesian network models for incomplete and dynamic data, *Statistica Neerlandica*, 74(3), pp. 397–419. doi: 10.1111/stan.12197.

Structural Expectation-Maximization (SEM)



Random Network Structure

Compute Expected sufficient statistics

A	B	C
1	??	34
26	8	19

Incomplete Data

E-Step

Replace

Update

Iterations

M-Step

Search for a BN structure that maximises the expected score function

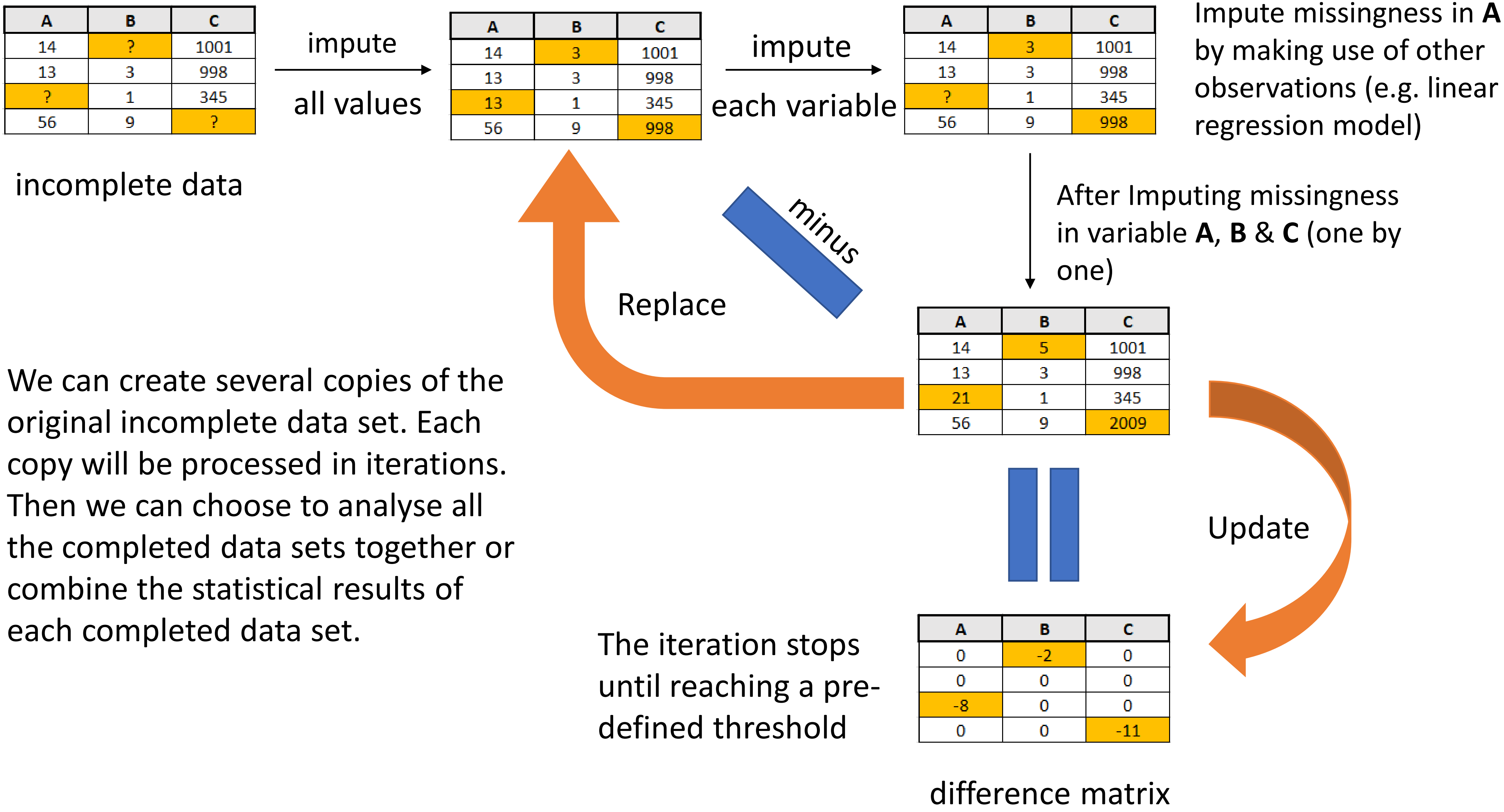
Converge

Update

A	B	C
1	77	34
26	8	19

Completed Data

Multiple Imputation by Chained Equations (MICE)



A	B	C
14	?	1001
13	3	998
?	1	345
56	9	?

incomplete data

impute

all values

A	B	C
14	3	1001
13	3	998
13	1	345
56	9	998

impute

each variable

A	B	C
14	3	1001
13	3	998
?	1	345
56	9	998

Impute missingness in **A** by making use of other observations (e.g. linear regression model)

After Imputing missingness in variable **A**, **B** & **C** (one by one)

A	B	C
14	5	1001
13	3	998
21	1	345
56	9	2009

minus

Replace

Update

A	B	C
0	-2	0
0	0	0
-8	0	0
0	0	-11

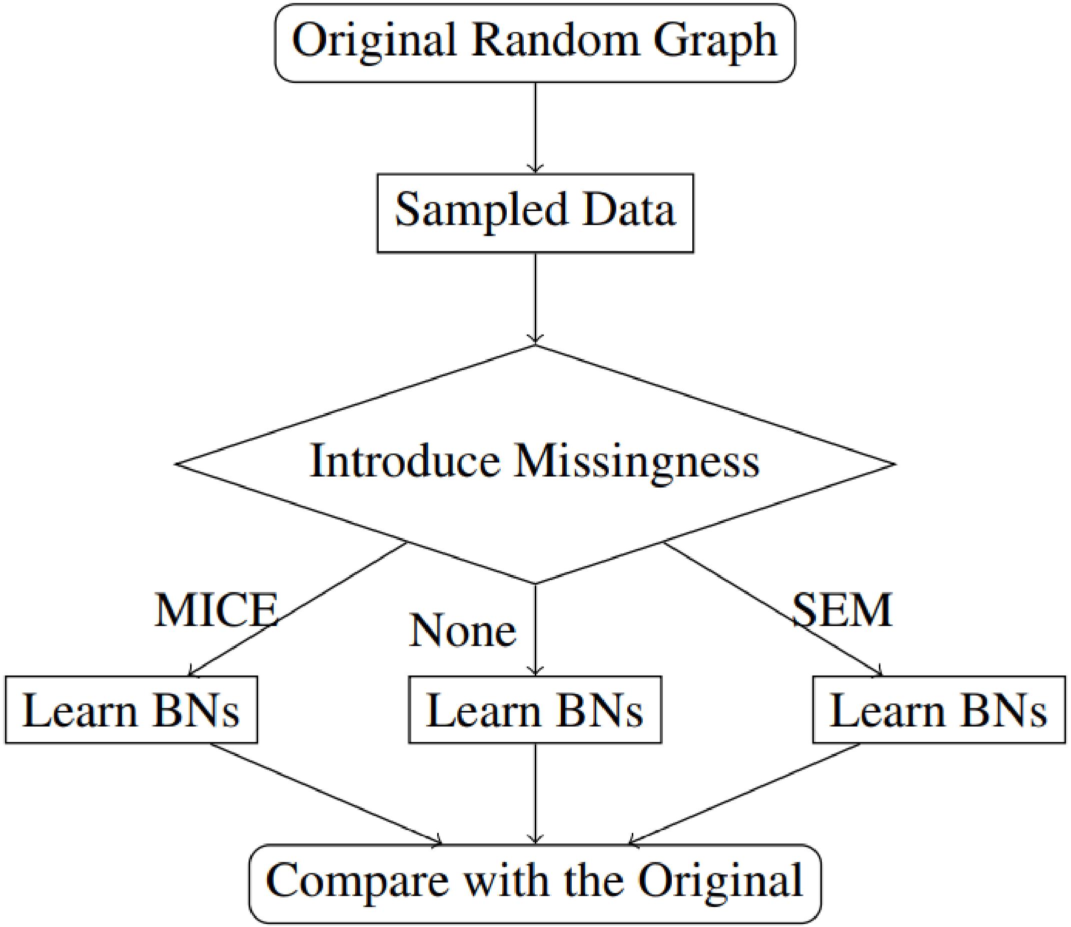
difference matrix

The iteration stops until reaching a pre-defined threshold

We can create several copies of the original incomplete data set. Each copy will be processed in iterations. Then we can choose to analyse all the completed data sets together or combine the statistical results of each completed data set.

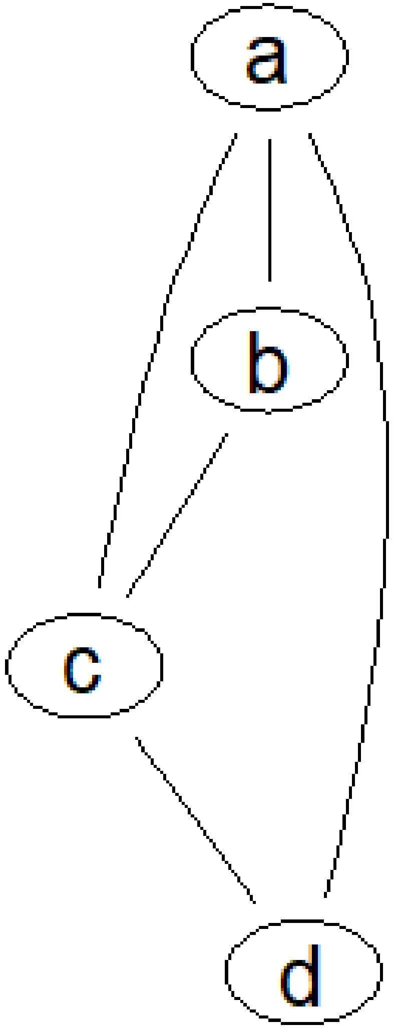
Compare the performance of SEM and MICE

Simulation Study

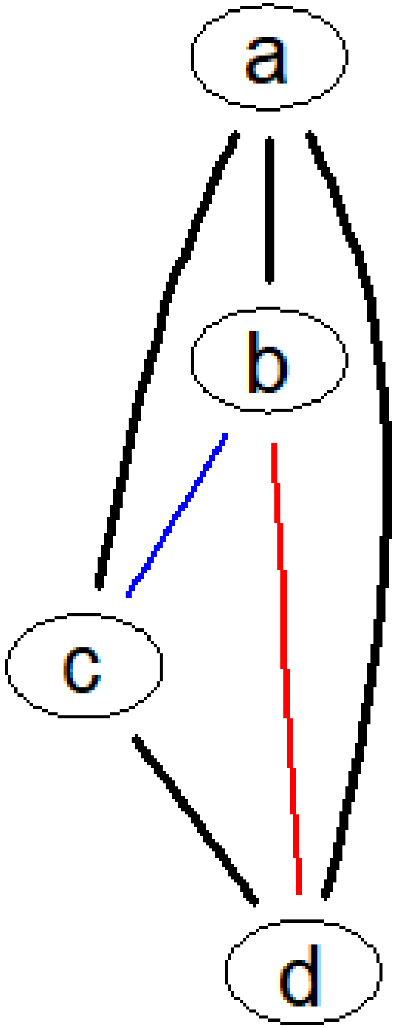


- Variables: 2 to 20
- Data points: 1000, 5000, 10000
- Missing proportion: 0.1 to 0.6 at intervals of 0.1
- Each condition is repeated 100 times.

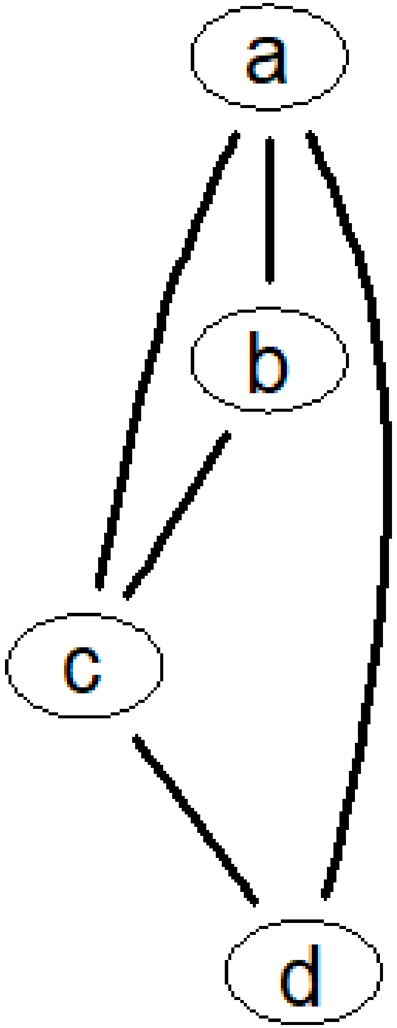
Original BN



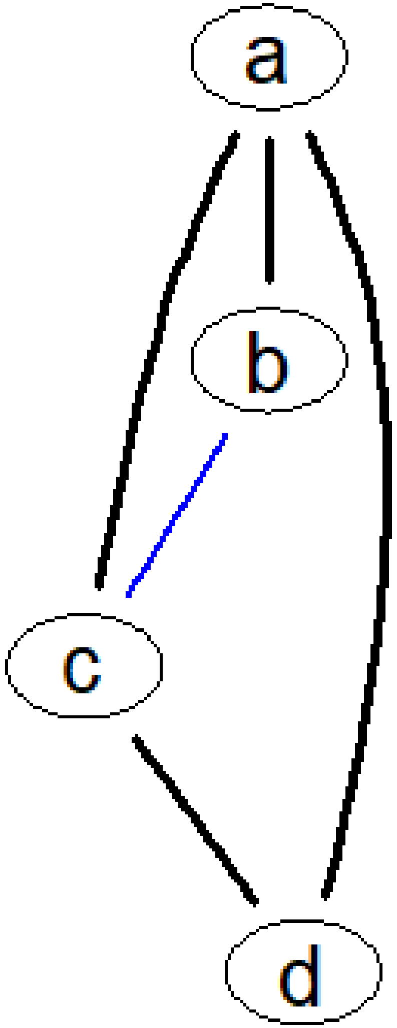
None

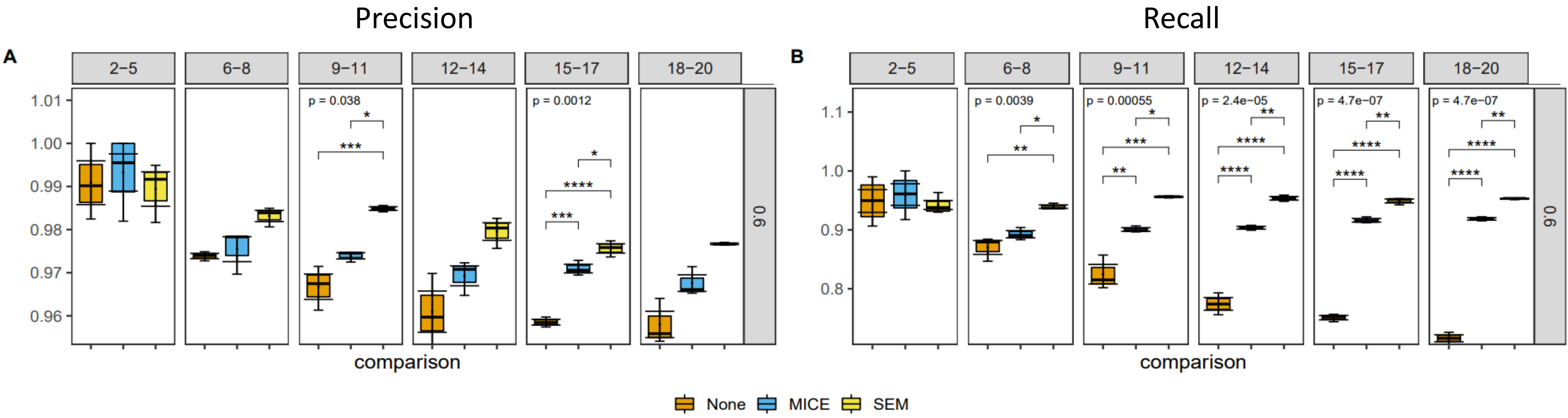


SEM



MICE



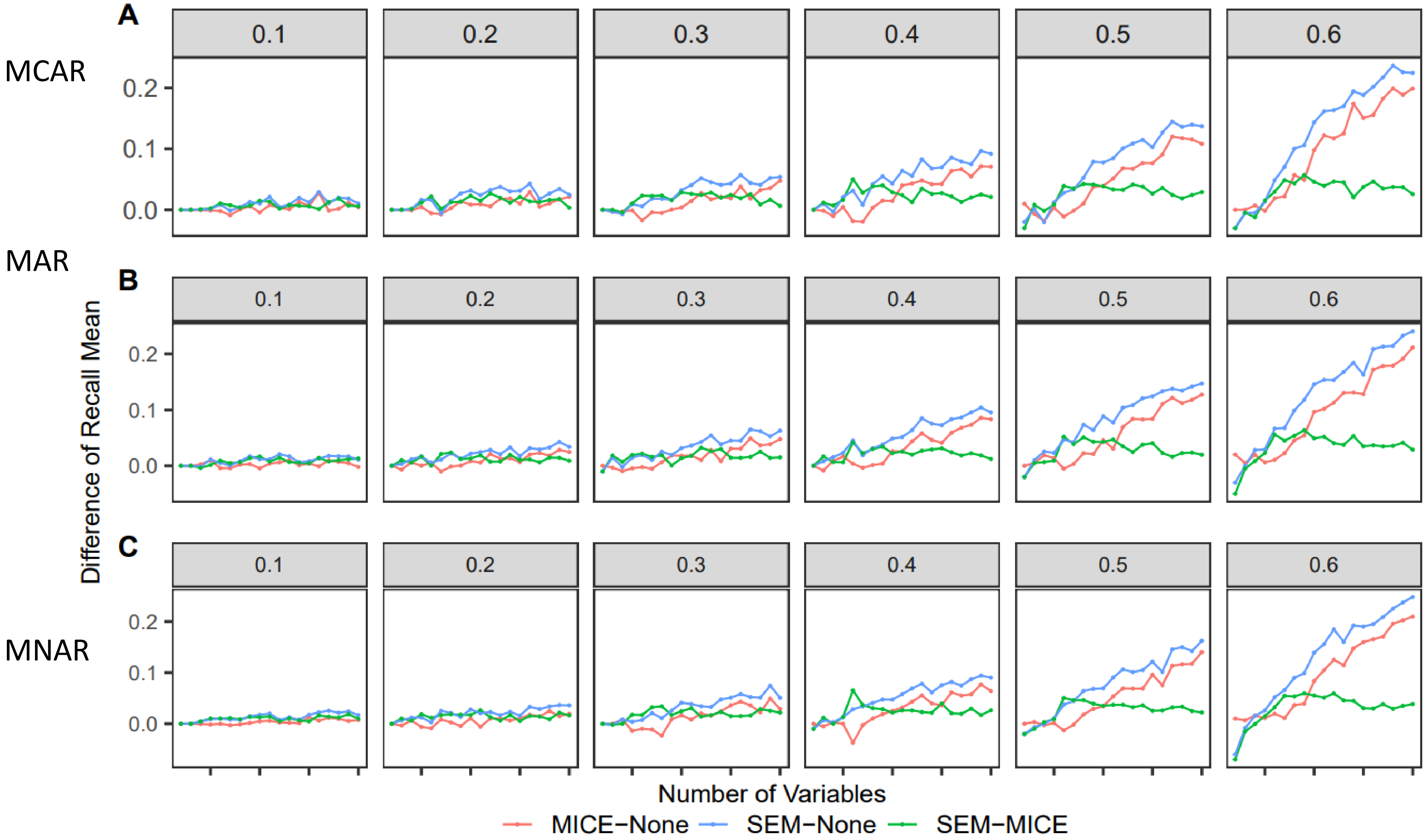


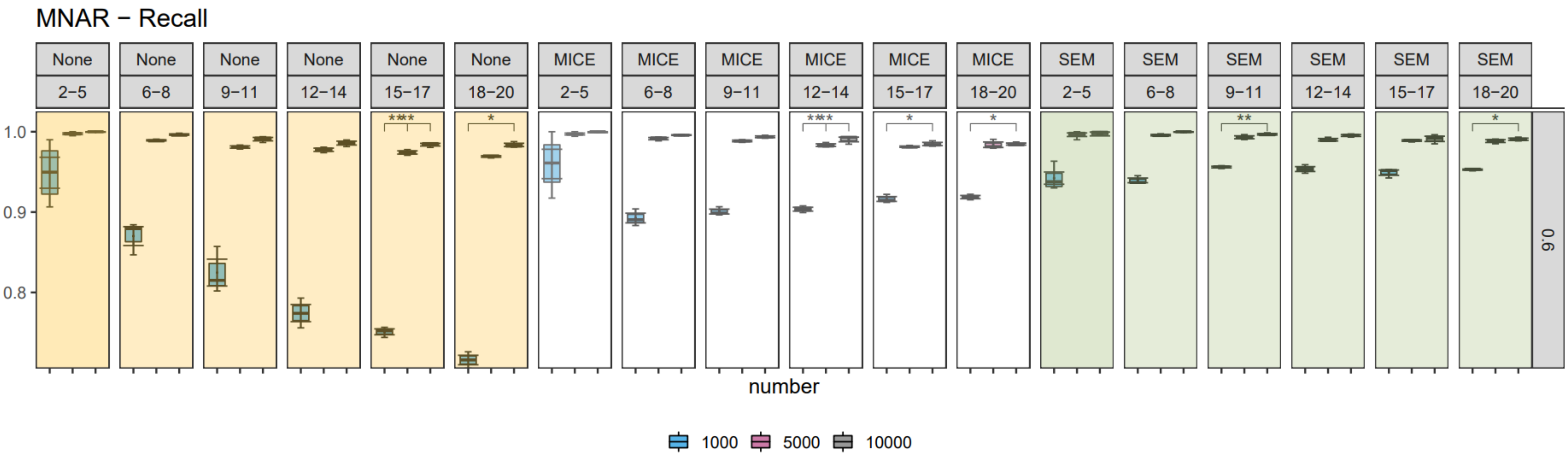
Data points: 1000

MNAR Data

$$\textbf{A.}\ \text{Precision} = \frac{TP}{TP+FP} \qquad \textbf{B.}\ \text{Recall} = \frac{TP}{TP+FN}$$

Statistical tests: One-way ANOVA, Tukey’s HSD pairwise tests, *, p < 0.05; **, p < 0.01; ***, p < 0.001; ****, p < 0.0001


A
MCAR
MAR
B
C
MNAR
0.1
0.2
0.3
0.4
0.5
0.6
0.0
0.1
0.2
Difference of Recall Mean
Number of Variables
MICE−None
SEM−None
SEM−MICE

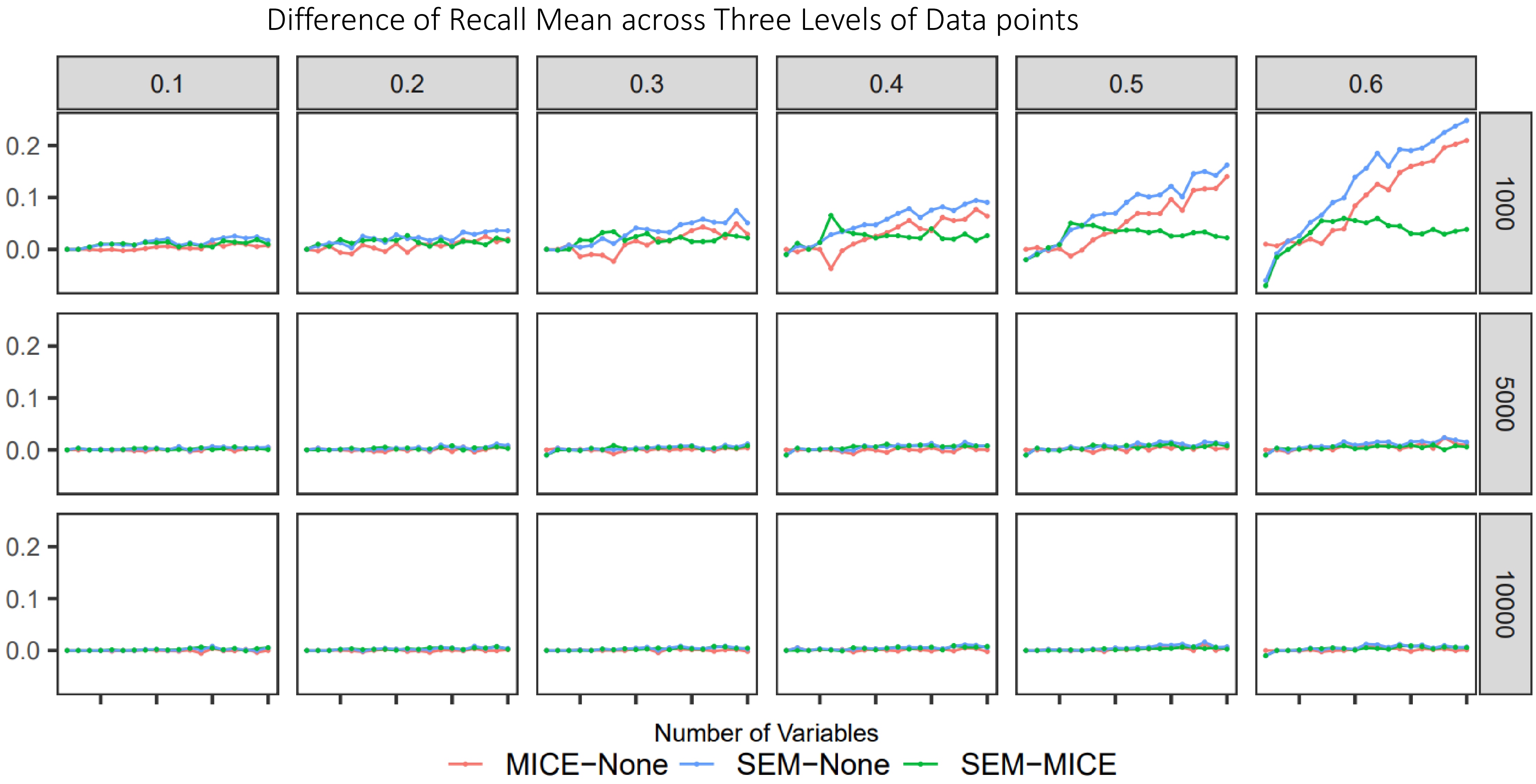


Comparison across three levels of data points.

MNAR Data

$$\text{Recall} = \frac{TP}{TP+FN}$$

Statistical tests: One-way ANOVA, Tukey’s HSD pairwise tests, *, p < 0.05; **, p < 0.01; ***, p < 0.001; ****, p < 0.0001


Difference of Recall Mean across Three Levels of Data points
0.1
0.2
0.3
0.4
0.5
0.6
1000
5000
10000
0.2
0.1
0.0
Number of Variables
MICE−None
SEM−None
SEM−MICE

MNAR Data

Conclusion

- Both SEM and MICE ↑ the completeness of Bayesian network structure learned from incomplete dataset.
- In some circumstances (e.g., data with high missing proportion and high number of variables), the performance of SEM algorithm > MICE.
- When there are low number of data points, the outperformance of SEM over the other two methods ↑ with ↑ number of variables and ↑ missing proportion.
- The outperformance of SEM and MICE over doing nothing decreases ↓ when there are high number of data points.

Case study on AMR data

University of
St Andrews
FOUNDED
1413

HATUA
Holistic Approach To
Unravel Antibacterial
Resistance in East Africa

UK Research
and Innovation

Holistic approach to unravel antibacterial resistance in East Africa (HATUA)

Variables (13)	Description	Levels
gender	Gender of each patient	"Male" , "Female"
age	Age of each patient	"<35", "35-64", "65 and above", NA
health_cost	How has it been for the patient to meet the cost of your own healthcare needs in the last 12 months?	"Very difficult", "little difficult", "Easy", NA
hospital_level	From which level of hospital has the patient been recruited?	"high", "low"
self_treatment	How did the patient first seek treatment?	"Non Self-treatment", "Self-treatment", NA
antibiotic_taking	What drugs did the patient take while seeking treatments?	"Yes antibiotic consumption", "No antibiotic consumption", NA
steps_pathway	The UTI pathway steps that patients took in seeking treatments.	"complex pathway: 2+ steps", "simple pathway: 0/1 step", NA
doctor_prescript	Did doctors give the patient a prescription (line) for antibiotics?	"no", "yes", NA
medicine_taking	What kind of medicines did the patient take for subsequent treatment?	"No medicine", "AB suitable for UTI", "Other AB", NA
see_doctor	Have the patient ever been to the doctor /hospital/health worker for these kinds of symptoms in the past?	"Yes", "No", NA
genus	The species that have been identified from the urine samples.	"Determined bacteria", "Undetermined bacteria", NA
gram_reaction	The gram reaction of species identified from the urine samples.	"negative", "positive", NA
MDR	Whether the patient has multiple frug resistance (MDR) infection.	"yes", "no", NA

“Multiple drug resistance (MDR), multidrug resistance or multi-resistance is AMR shown by a species of microorganism to at least one antimicrobial drug in three or more antimicrobial categories. ”

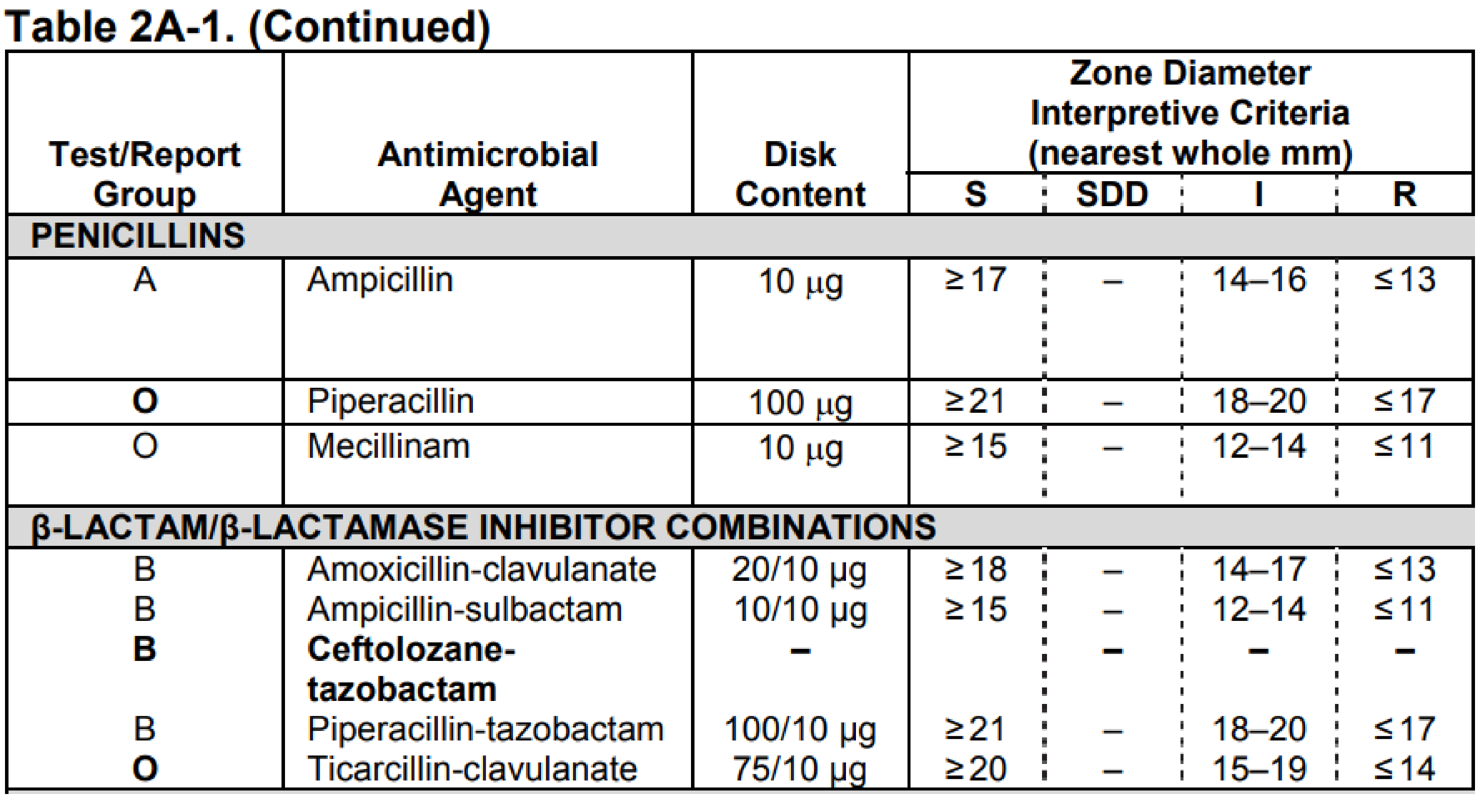
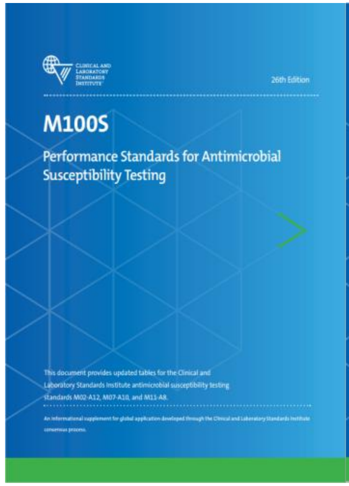


Table 2A-1. (Continued)

Test/Report Group	Antimicrobial Agent	Disk Content	Zone Diameter Interpretive Criteria (nearest whole mm)			
			S	SDD	I	R
PENICILLINS						
A	Ampicillin	10 µg	≥17	–	14–16	≤13
O	Piperacillin	100 µg	≥21	–	18–20	≤17
O	Mecillinam	10 µg	≥15	–	12–14	≤11
β-LACTAM/β-LACTAMASE INHIBITOR COMBINATIONS						
B	Amoxicillin-clavulanate	20/10 µg	≥18	–	14–17	≤13
B	Ampicillin-sulbactam	10/10 µg	≥15	–	12–14	≤11
B	**Ceftolozane-tazobactam**	**–**		**–**	**–**	**–**
B	Piperacillin-tazobactam	100/10 µg	≥21	–	18–20	≤17
O	Ticarcillin-clavulanate	75/10 µg	≥20	–	15–19	≤14



Clinical and Laboratory Standards Institute

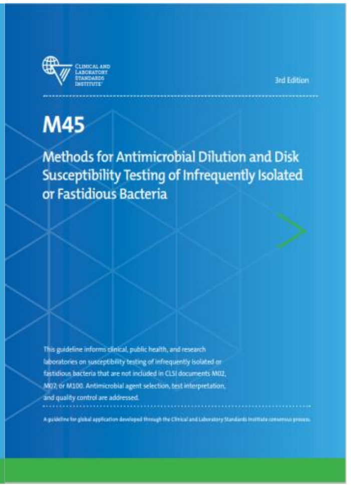
26th Edition

M100S

Performance Standards for Antimicrobial Susceptibility Testing

This document provides updated tables for the Clinical and Laboratory Standards Institute antimicrobial susceptibility testing standards M02-A12, M07-A10, and M11-A8.

An international supplement for global application developed through the Clinical and Laboratory Standards Institute consensus process.



Clinical and Laboratory Standards Institute

3rd Edition

M45

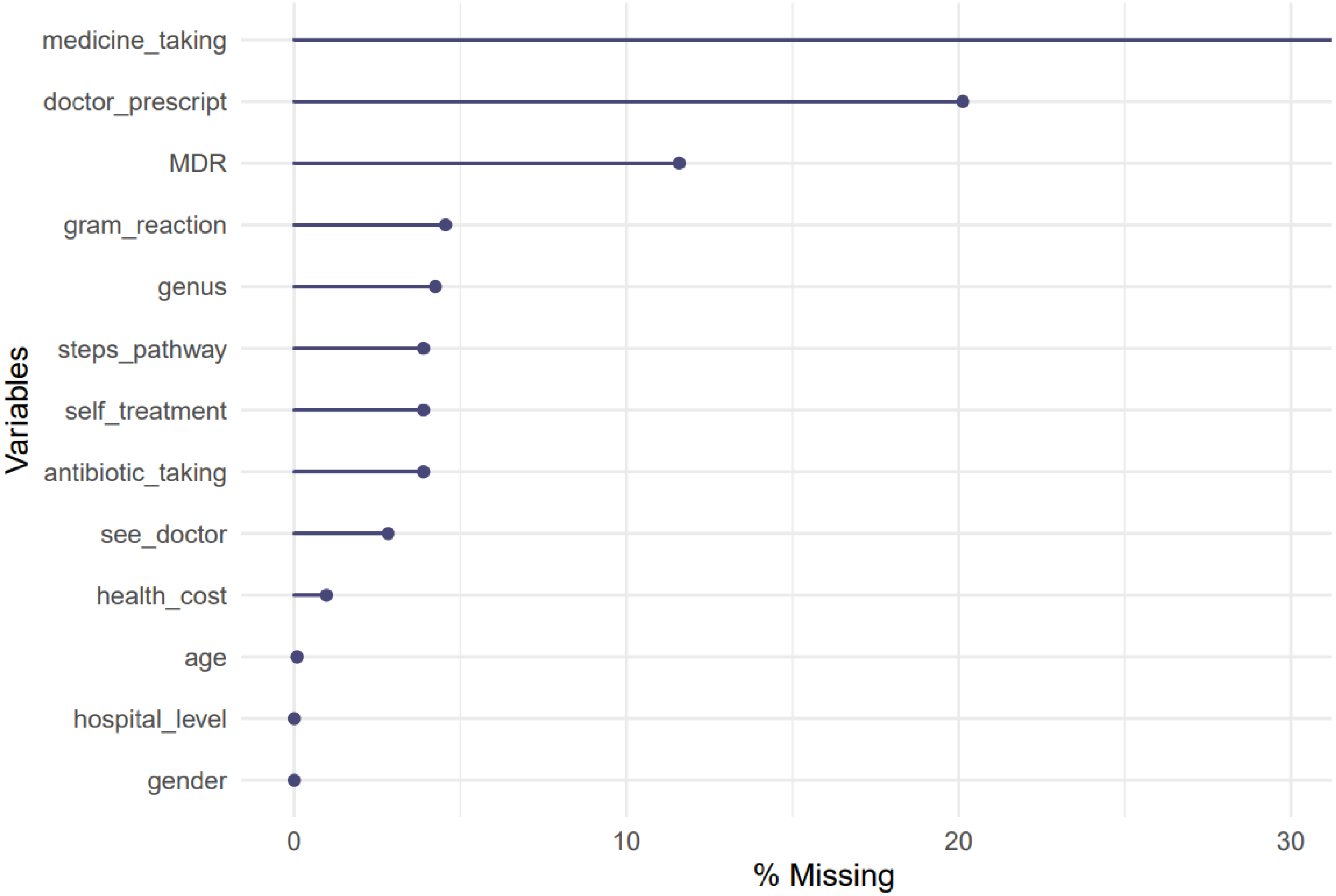
Methods for Antimicrobial Dilution and Disk Susceptibility Testing of Infrequently Isolated or Fastidious Bacteria

This guideline informs clinical, public health, and research laboratories on susceptibility testing of infrequently isolated or fastidious bacteria that are not included in CLSI documents M02, M07, or M100. Antimicrobial agent selection, test interpretation, and quality control are addressed.

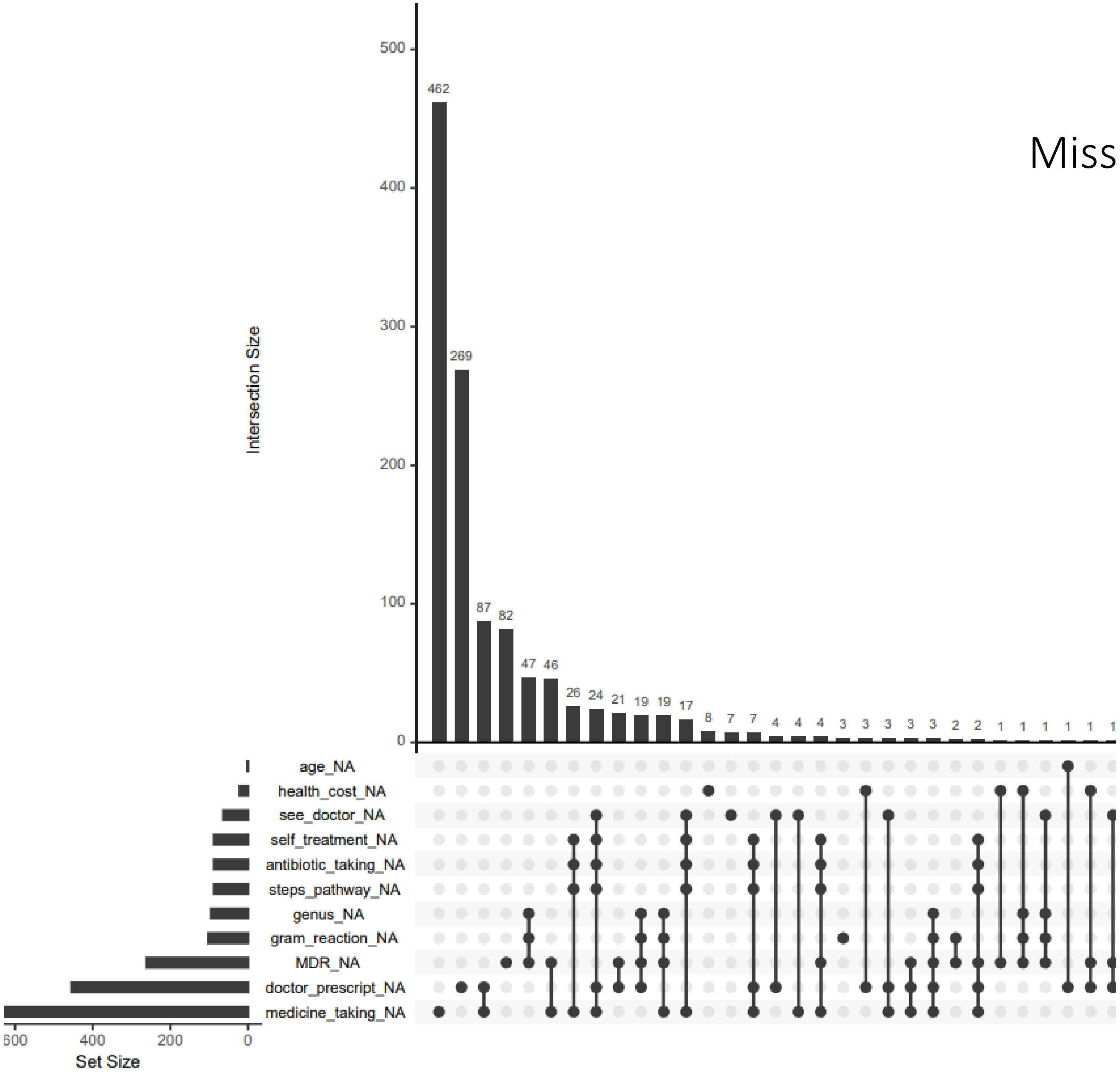
A guideline for global application developed through the Clinical and Laboratory Standards Institute consensus process.

Magiorakos AP, Srinivasan A, Carey RB, Carmeli Y, Falagas ME, Giske CG, Harbarth S, Hindler JF, Kahlmeter G, Olsson-Liljequist B, Paterson DL, Rice LB, Stelling J, Struelens MJ, Vatopoulos A, Weber JT, Monnet DL. Multidrug-resistant, extensively drug-resistant and pandrug-resistant bacteria: an international expert proposal for interim standard definitions for acquired resistance. Clin Microbiol Infect. 2012 Mar;18(3):268-81. doi: 10.1111/j.1469-0691.2011.03570.x. Epub 2011 Jul 27. PMID: 21793988.
https://onlinelibrary.wiley.com/doi/pdf/10.1111/j.1469-0691.2011.03570.x

Distribution of Missing Values



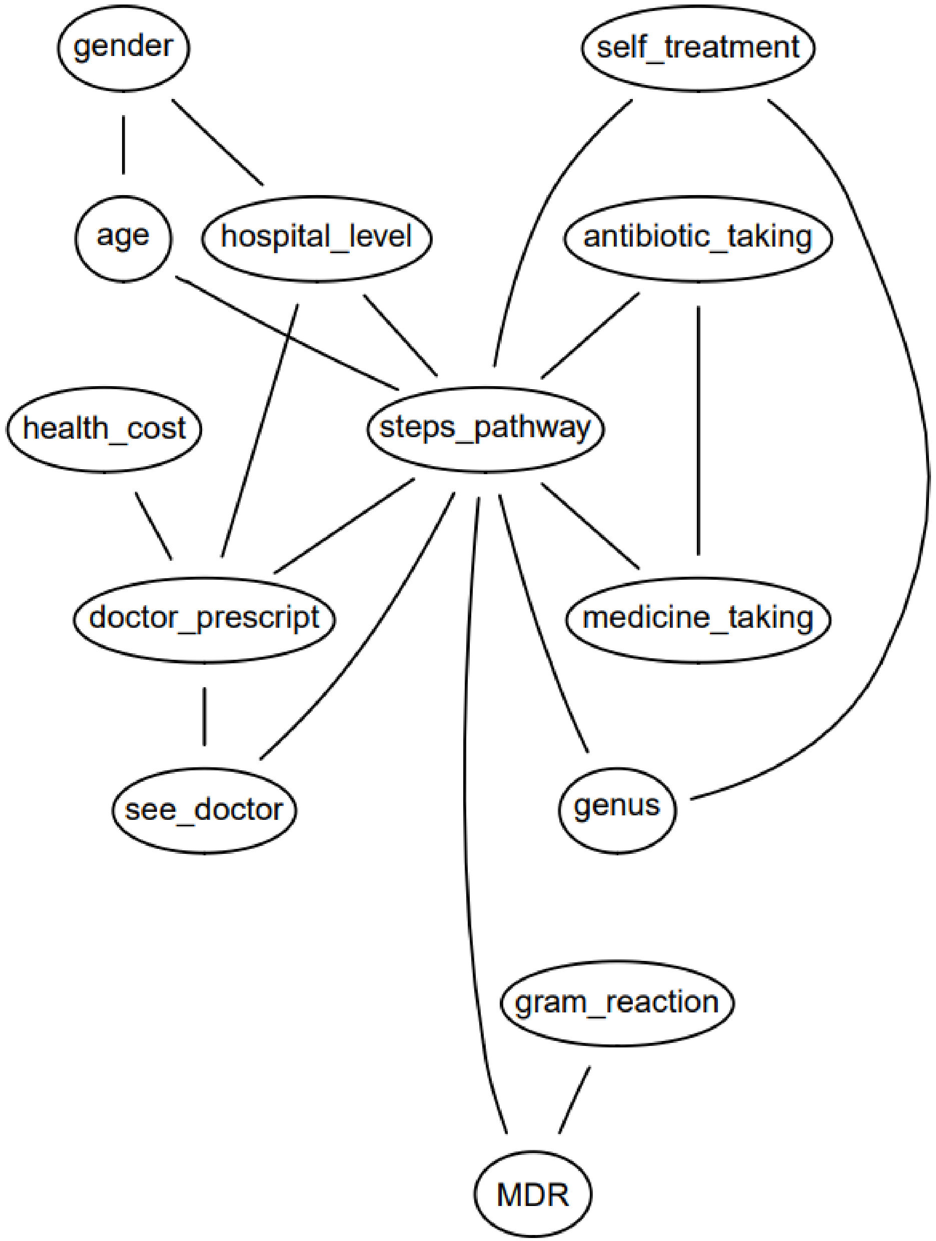
Missing Patterns



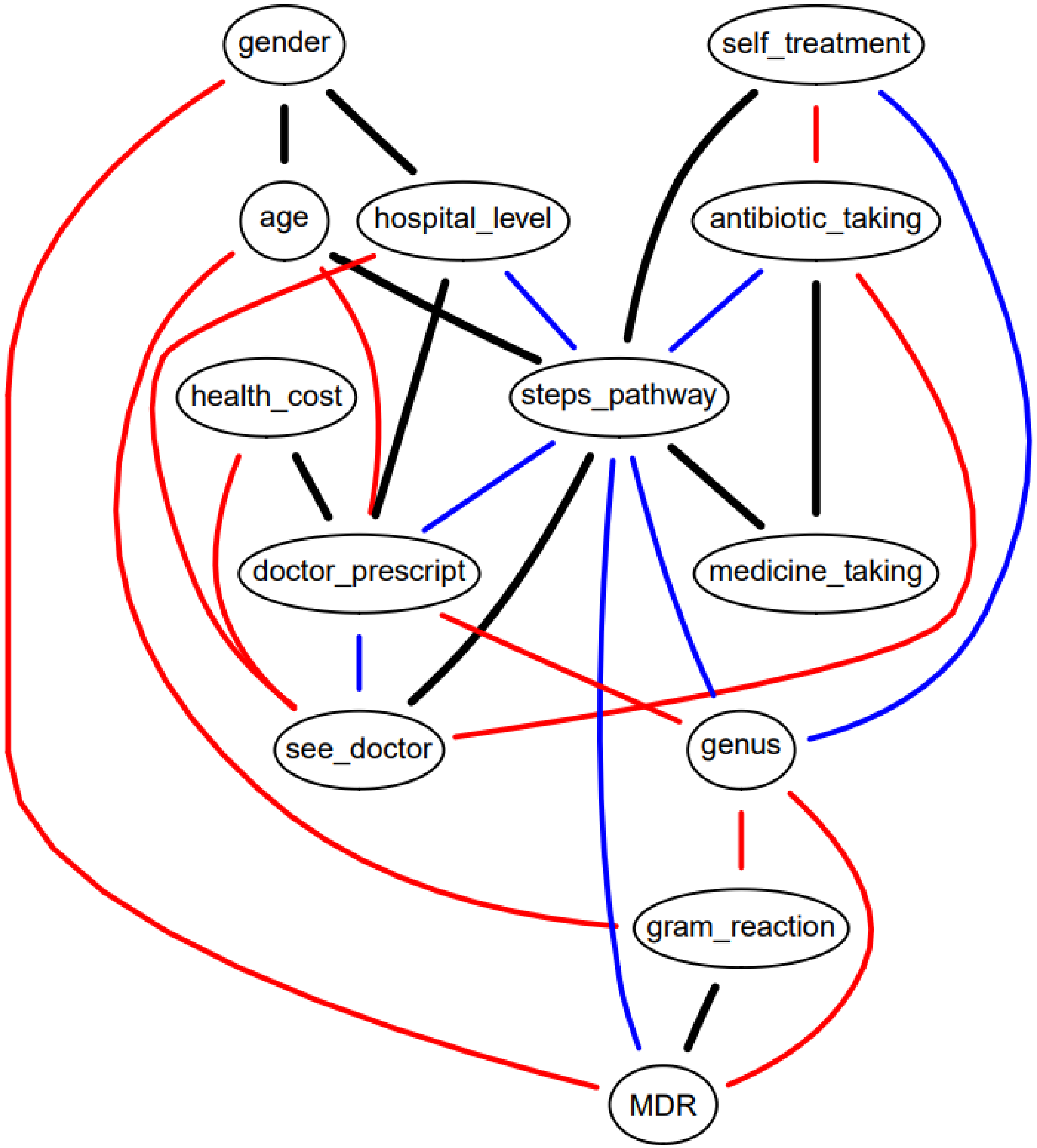
- Observations N = 2261
- Complete cases N = 1067

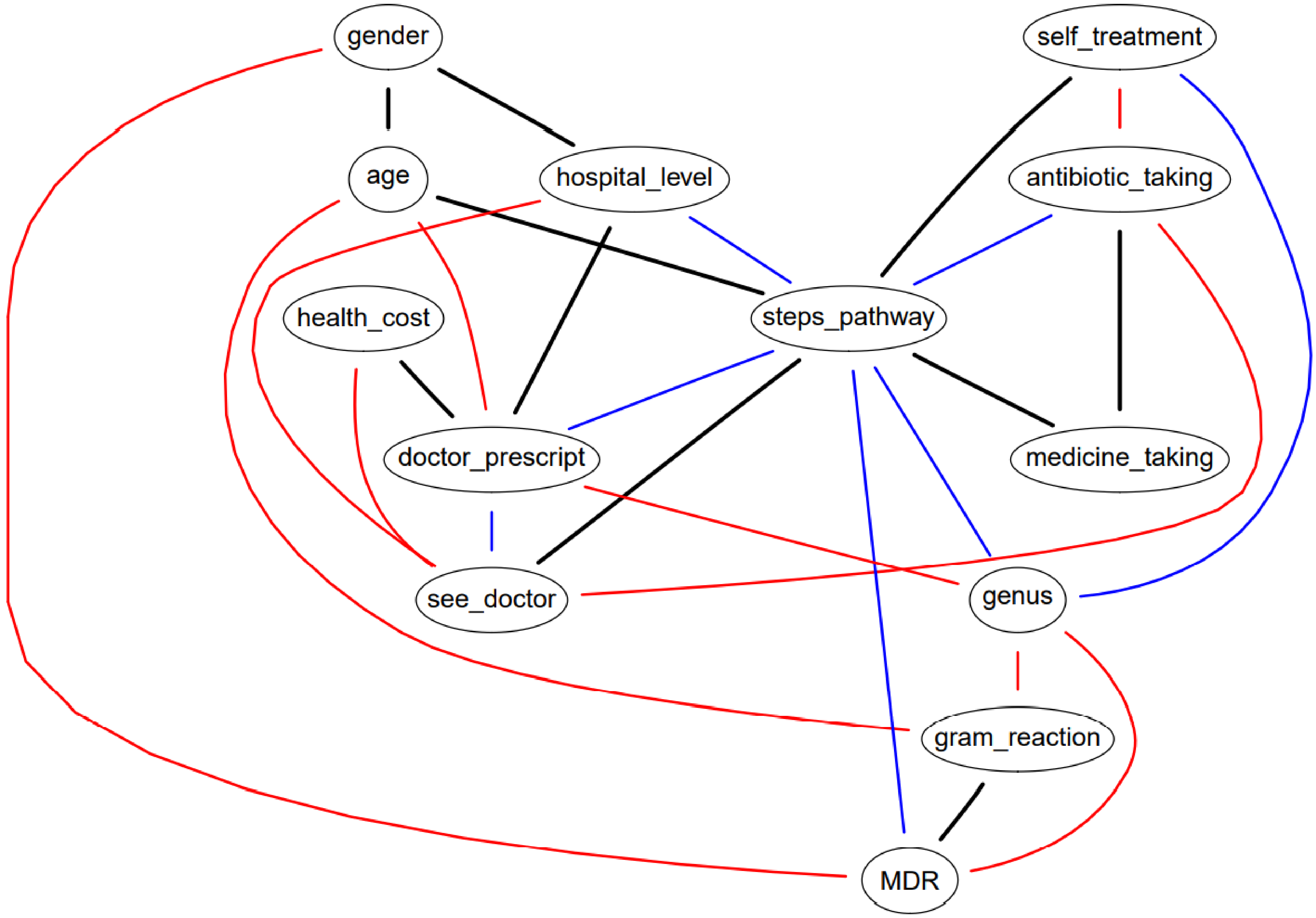
We checked variables that are missing together and re-coded some missing values with known reasons of missingness.

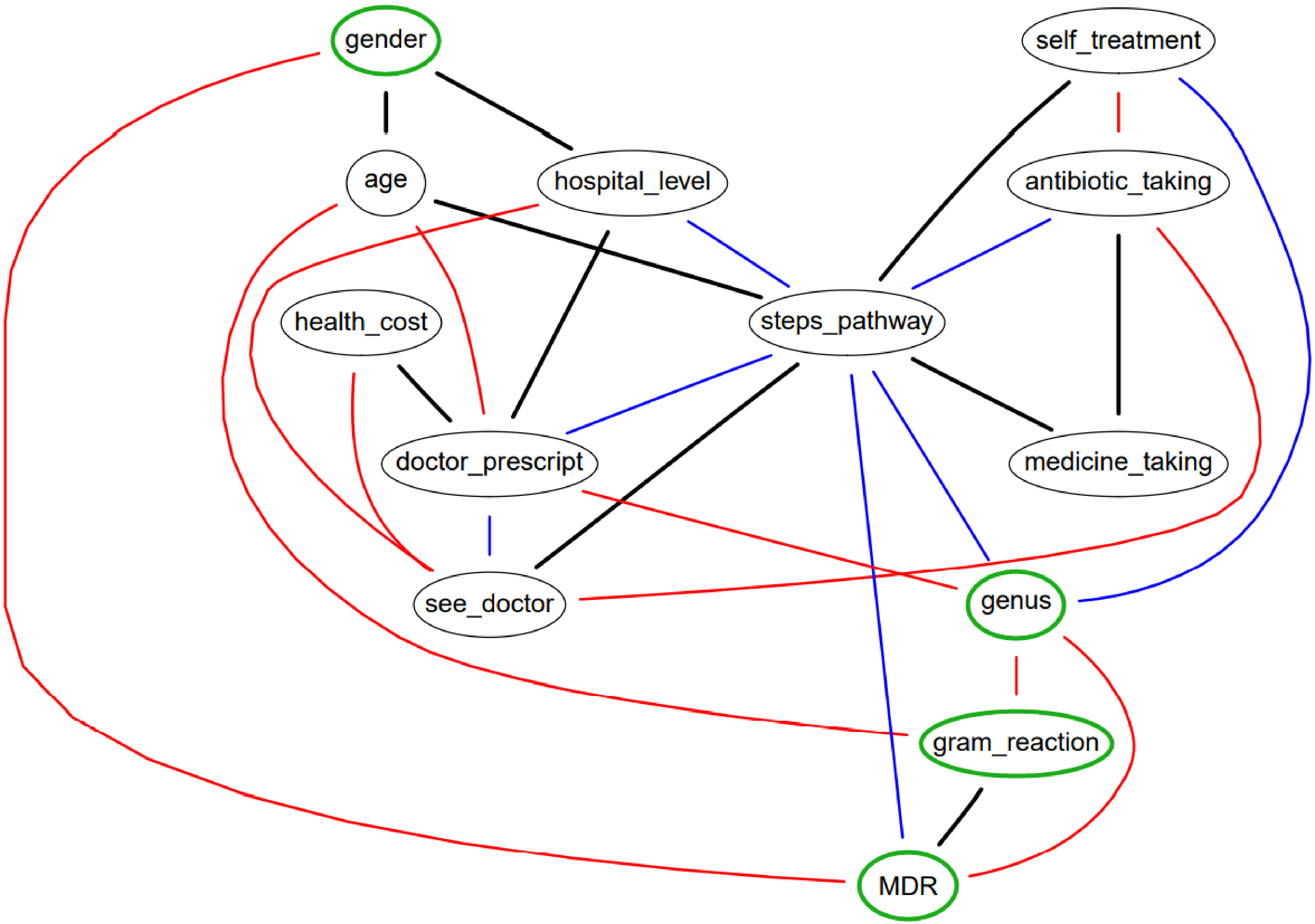
Complete cases

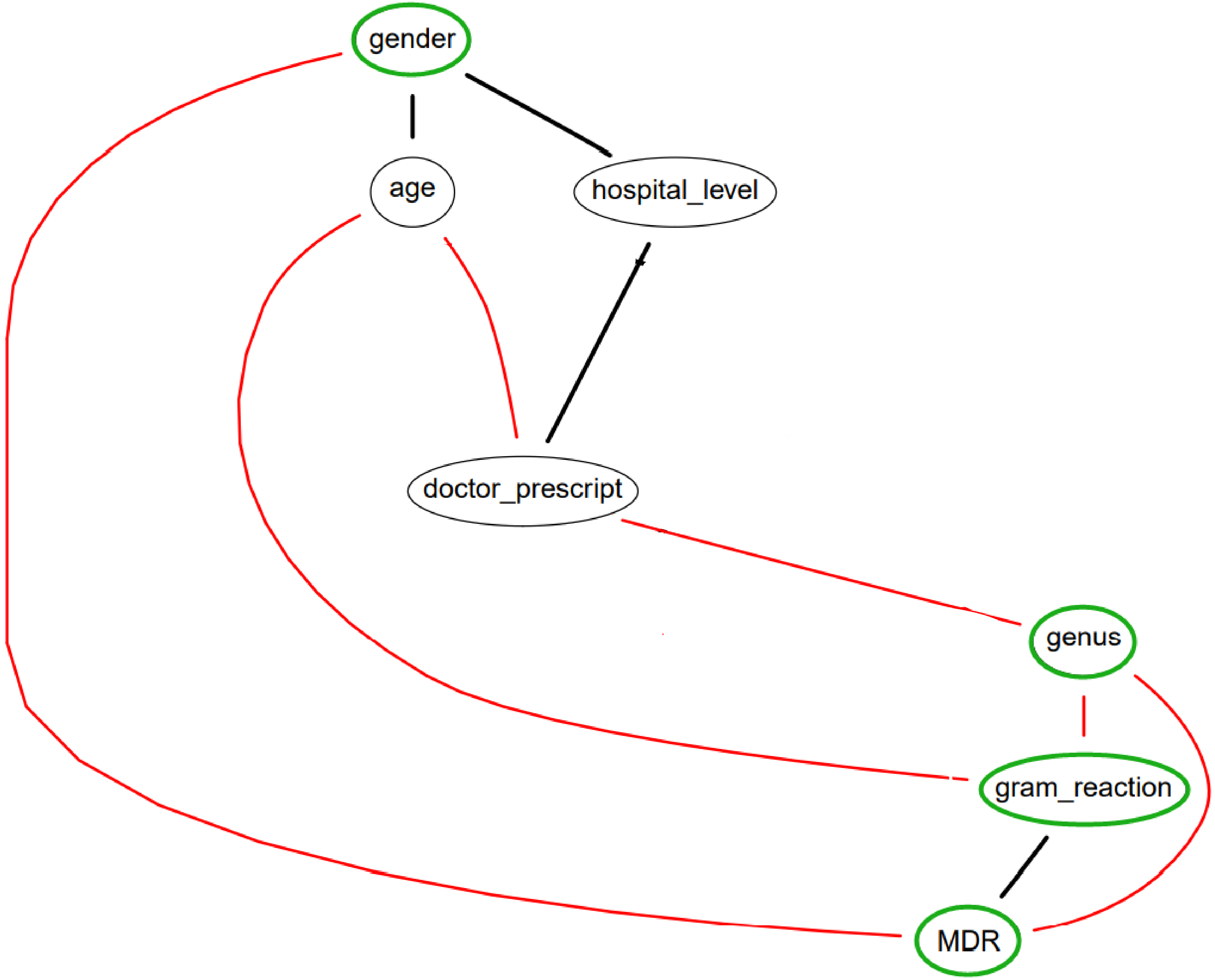


SEM




gender
self_treatment
age
hospital_level
antibiotic_taking
health_cost
steps_pathway
doctor_prescript
medicine_taking
see_doctor
genus
gram_reaction
MDR


gender
self_treatment
age
hospital_level
antibiotic_taking
health_cost
steps_pathway
doctor_prescript
medicine_taking
see_doctor
genus
gram_reaction
MDR


gender
age
hospital_level
doctor_prescript
genus
gram_reaction
MDR

Conclusion

- SEM algorithm identified factors associated with multiple-drug resistance
  - Genus of bacteria that infected patients
  - Type of bacteria (gram - positive or negative)
  - Gender of patients
  - Patient's age
  - Whether patients have been provided the prescription of antibiotics from doctors
  - The hospital level that patients have been to seek treatments

To be continued…

Acknowledgement

- Dr V Anne Smith
- Dr Katherine Keenan

UK Research and Innovation

University of St Andrews

Scottish Graduate School of Social Science

HATUA

Holistic Approach To Unravel Antibacterial Resistance in East Africa

THANKS FOR LISTENING!

ANY QUESTIONS?