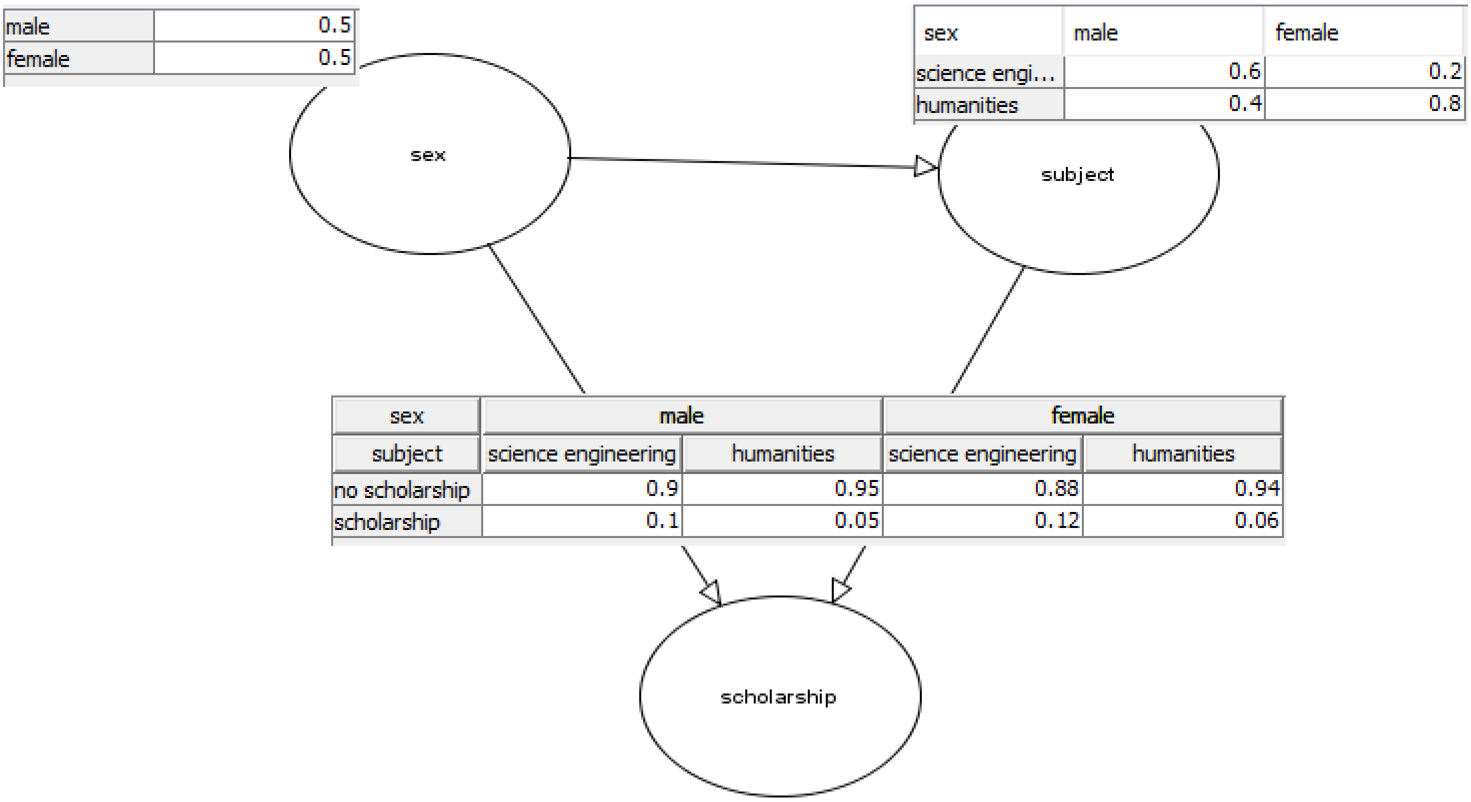
ECS7005P Elliot Linsey Summer 2022

Question 1:

a) C x 0.02 + (1-C) x 0.01
b) 0.3 x 0.02 + 0.7 x 0.01 = 0.013 or 1.3%
c) 0.02 x t + 0.98 x f
d) (0.02 x t)/(0.02 x t + 0.98 x f)
e) 0.02 x 0.8 + 0.98 x 0.1 = 0.114 or 11.4%
f) 0.02 x 0.8/0.114 = 0.140 or 14%
g) 0.01 x 0.9 + 0.99 x 0.15 = 0.1575 or 15.75%
h) 0.01 x 0.9/0.1575 = 0.0571 or 5.7%
i) 0.114 x 0.3 + 0.1575 x 0.7 = 0.144 or 14.4%
j) 0.02 x 0.2/(0.02 x 0.2 + 0.98 x 0.9) = 0.00451 or 0.45%
k) 0.01 x 0.1/(0.01 x 0.1 + 0.99 x 0.85) = 0.00119 or 0.119%
l) 0.14 x 0.8 + 0.86 x 0.1 = 0.19 or 19%

Question 2:

a) 76/1000 = 0.076 or 7.6%
b) 42/400 = 0.105 or 10.5%
c) 34/600 = 0.057 or 5.7%
d) 30/300 = 0.1 or 10%
e) 10/200 = 0.05 or 5%
f) 12/100 = 0.12 or 12%
g) 24/400 = 0.06 or 6%
h) That overall, males were awarded the most scholarships with 8% and females achieving 7.2%, but when drilled down we find that females got higher numbers of scholarships in both subjects.
i) Simpson’s Paradox
j) The main cause is sex being a confounding variable. Females were more likely to apply for humanities and less science and engineering with males being the other way around.



male	0.5
female	0.5

sex	male	female
science engi…	0.6	0.2
humanities	0.4	0.8

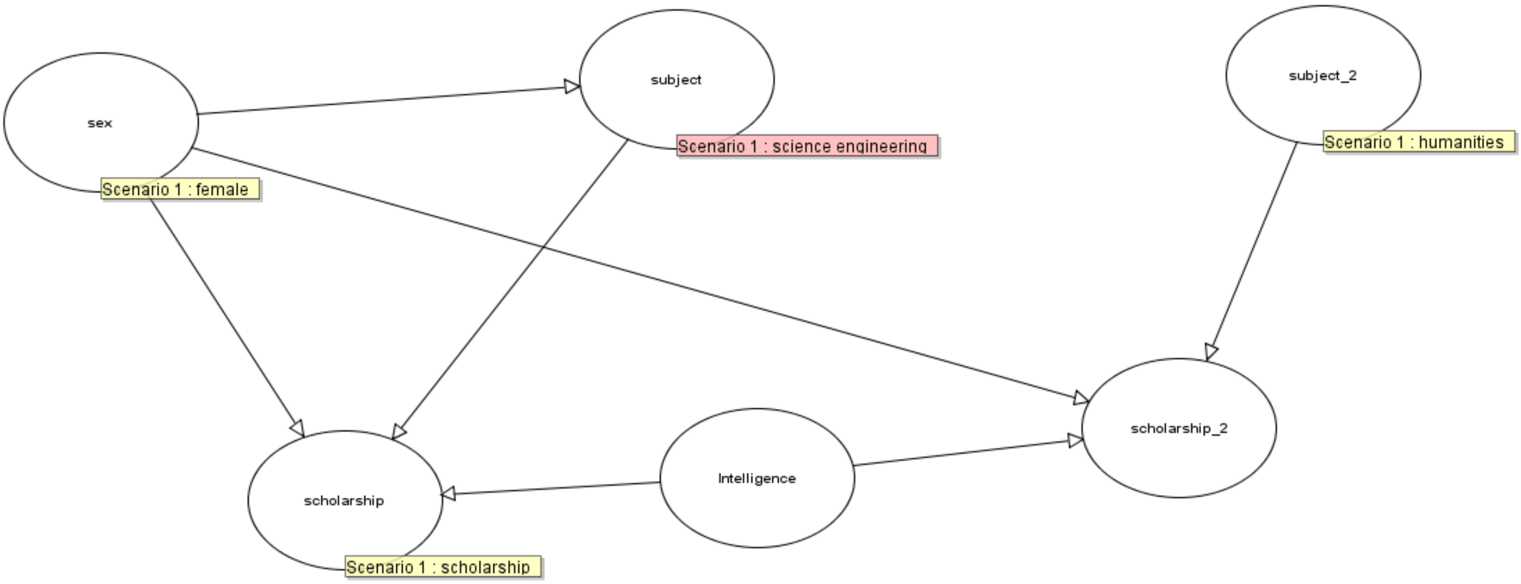
sex	male		female	
subject	science engineering	humanities	science engineering	humanities
no scholarship	0.9	0.95	0.88	0.94
scholarship	0.1	0.05	0.12	0.06

k)

l) Cut the link between sex and subject

m) Male = 7.5% and female = 9%

n) We first put our intelligence node that causally links to scholarship. In the real world we set our nodes to 'female', 'science and engineering' and 'scholarship'. We then copy the scholarship and subject nodes to a counterfactual world, cut the link between sex and subject again and set subject to 'humanities' and observe the probability for her obtaining a scholarship.



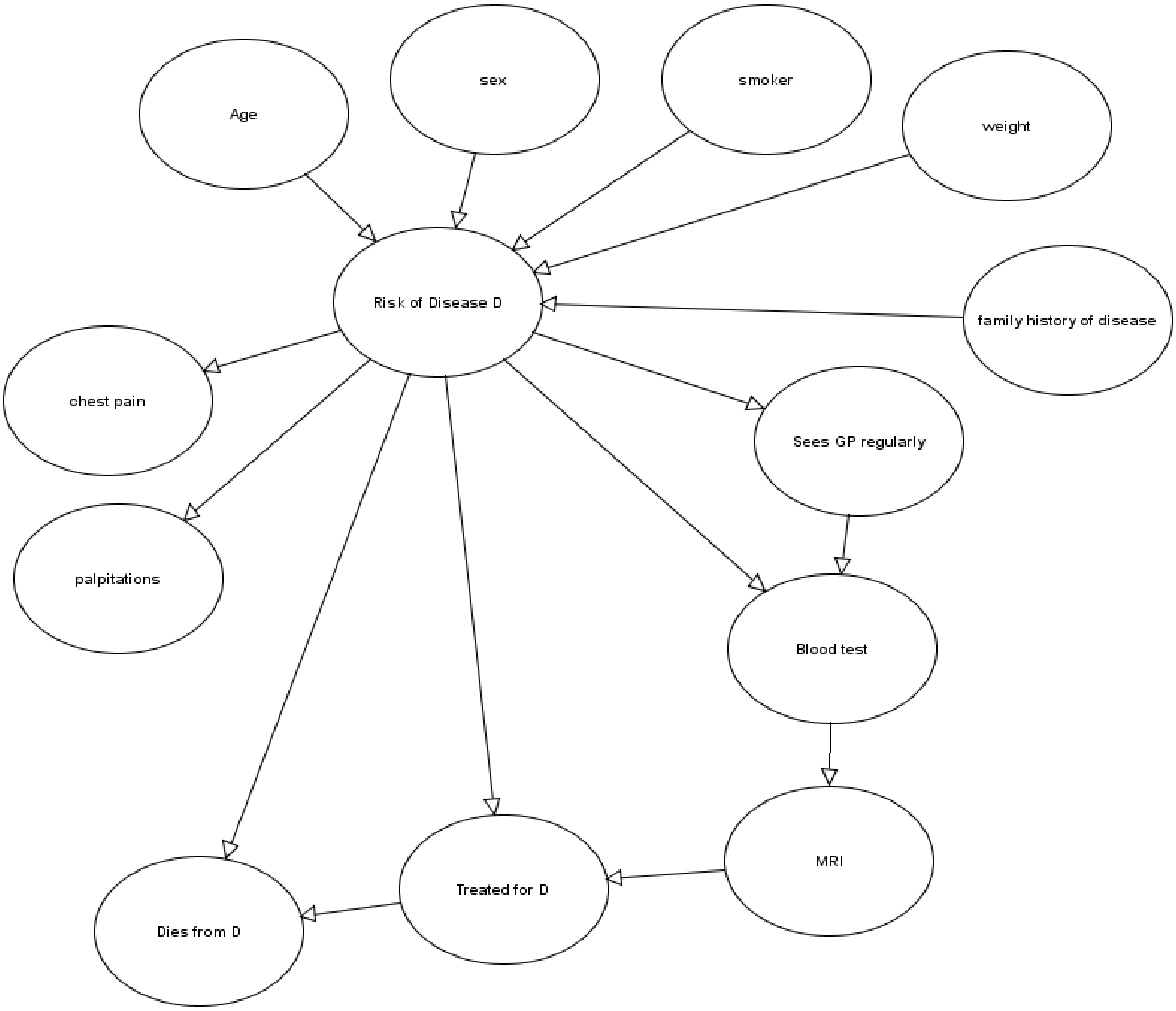
Question 3:

a) 140/7000 = 0.02
b) 90/3000 = 0.03
c) 0.03-0.02/0.02 = 0.5 or 50%
d) 0.03-0.02 = 0.01 or 1%
e) 0.03/0.02 = 1.5
f) 0.3 x 0.03/(0.3 x 0.03 + 0.7 x 0.02) = 39.13%
g) This is prosecutor's fallacy, they have mistaken P(cancer|non-smoker) with P(non-smoker|cancer).

Question 4:

a) The model would be logistic regression as the outcome is binary of 'survived', 'died'. It would take the features and predict a probability of survival ranging from 0 to 1. Depending on a threshold set, any probability over this threshold would be classed as 'survived'. These results would be compared to the actual results and the accuracy is calculated by summing the number of correct classifications over the total number of classifications
b) The dataset lacks the feature of family history of the disease. Values are too coarse; smoker could have length of time being a smoker and chest pain could have severity of chest pain. Missing variables, it says it means 'not recorded' but does this mean the test was not done? Is it possible to get this data? the more data the more accurate the prediction. Does not tell us whether the person was treated for the disease, this is extremely important piece of data and without it the dataset is essentially censored. (Question C is on the next page)

c)


Age
sex
smoker
weight
Risk of Disease D
family history of disease
chest pain
Sees GP regularly
palpitations
Blood test
MRI
Treated for D
Dies from D