Case Study

A group of population health researchers at a major urban hospital are interested in quantifying the long-term risks of heart disease. The researchers are interested in predicting outcomes over the next ten years. The researchers seek to provide practical advice for general practitioners (GPs), who must screen patients for further diagnosis and treatment. The GP is the person that provides the first line of care in the support of the patient.

Questions

1. Over a ten year period the researchers notice four clinically significant outcomes. The first is no heart disease. The second outcome is distress, not resulting in hospitalization. The third is hospitalization for heart symptoms regardless of other distress. The fourth possible event is the death of the patient by heart attack. (This question is worth 10 points out of 100.)

   a. Create a probability mass function for the four events. You may choose any consistent probabilities of your choice (5/10).
   b. Discuss the challenges involved in attaching probabilities to this distribution if some fraction of the patients in the records died from other causes than heart disease. A short answer is expected here (5/10).

1a

E_1	$p_1 = 0.40$
E_2	$p_2 = 0.30$
E_3	$p_3 = 0.20$
E_4	$p_4 = 0.10$

Event probabilities must sum to 1.00.

1b. The events may not span the space of possible events. The events may not be independent. These are required of probabilities

Restating that you are only studying heart attack is a given. Reassuring credibility is wrong.

1a 1pt/s for algebra but no quantities

1b 2pt/s for pragmatic answer
3pt/s for policy relevant answer

Compare prep question 5, 10,

2. A simplified screening model involves cholesterol level and age. Cholesterol level comes in three levels – low, medium and high. Age comes in three levels – young, middle-aged, and old. In this population 45% of the citizens have low cholesterol, 35% have medium cholesterol, and 20% have high cholesterol. In this population 15% of the population is young, 45% of the population is middle-aged, and 40% is old. Complete the table below by filling in the joint probabilities assuming that age and cholesterol level are independent of one another. (This question is with 10 points out of 100.)

	Young	Middle-Aged	Old	Totals
Low	0.15 · 0.45	0.45 · 0.45	0.40 · 0.45	0.45
Medium	0.15 · 0.35	0.45 · 0.35	0.40 · 0.35	0.35
High	0.15 · 0.20	0.45 · 0.20	0.40 · 0.20	0.20
Totals	0.15	0.45	0.40	1.00

Table 1. Age and Cholesterol, Marginal Distributions

	young	M-A.	old	totals	
low	0.0675	0.2025	0.1800	~~totals~~	0.45
medium	0.0525	0.1575	0.1400	~~0.45~~	0.35
high	0.0300	0.0900	0.0800	0.20	
totals	0.15	0.45	0.40	1.00	

I found no errors on this question.

Compare prep question 4, 6, 7

3. The actual joint distribution is given below. Are age and cholesterol independent of one another? How do you know? (This question is worth 10 points out of 100.)

	Young	Middle-Aged	Old	Totals
Low	0.10	0.25	0.10	0.45
Medium	0.05	0.20	0.10	0.35
High	0.00	0.00	0.20	0.20
Totals	0.15	0.45	0.40	1.00

Table 2. Age and Cholesterol, Joint Distribution

If any one of the marginals are not equal ~~independent~~ then the distributions are not independent

$p_1 = p(\text{low, young}) = 0.10$
(evidence)

$p_2 = p(\text{low, young}) = 0.15 \cdot 0.45 = 0.0675$
(marginals)

$p_1 \neq p_2$

The distributions are not conditionally independent

A discussion of independence was needed here to justify your answer

compare prep question 7

4. The population health specialists have made a model. They're not convinced it's the correct model, but they have collected evidence based on ten years of health records. In the table below the columns show the presence and absence of heart disease. In this case with D1 the disease is present, and D2 the disease is not present. The rows show the classification by the model. In this case M1 says the disease is present, and M2 says the disease is not present. (This problem is worth 15 points out of 100.)

	D1	D2	Total
M1	0.70	0.15	0.85
M2	0.10	0.05	0.15
Total	0.80	0.20	1.00

Table 3. The Probability of Disease Under the Model

a. Calculate the probability of the disease being present given the fact that the model says it is present. This quantity is p (M1 | D1). (5/15)
b. Calculate the probability that the data indicates the disease is present given the fact that the model says it is present. This quantity is p(D1 | M1). (5/15)
c. Are these two quantities the same? Why or why not? (5/15)

a $p(M1 \mid D1) = p(M1, D1) \mid p(D1) = 0.70 / 0.80 = 0,875$

b $p(D1 \mid M1) = p(M1, D1) \mid p(M1) = 0.70 / 0.85 = 0,824$

d The quantities are not the same since the marginal quantities are not identical $p(D1) \neq p(M1)$. This is the case in general.

Conditional probability calculations are expected here. If you did use Bayes' rule correctly you certainly received credit for the later question.

Compare prep questions 4, 6, 7.

5. The analysts would like to evaluate the credibility of their model. They would like to calculate the probability of the model given the data, or p(M1 | D1). What they have instead is the probability of the data given the model, or p(D1 | M1). (This problem is worth 10 out of 100.)

   a. Use the definition of conditional probabilities to relate these two quantities algebraically. Your answer should be an algebraic equation (5/10)
   b. Demonstrate algebraically the full evidence for p (D1). Expand this out so that you have the probability of D1 given M1 (disease presence given the model diagnosis), and the probability of D1 given M2 (disease presence given the lack of a model diagnosis). Your answer should be an algebraic equation (5/10).

a $p(M1 \mid D1) \cdot p(D1) = p(D1 \mid M1) \cdot p(M1)$

b $p(D1) = p(D1 \mid M2) + p(D1 \mid M1)$

Compare theory questions chapter 5

6. As given in the problem the analysts would like to calculate p(M1 | D1). (This question is worth 20 out of 100.)

   a. Write out the quantity which solves for p(M1 | D1) using Bayes' rule (5/20)?
   b. Given this, what algebraic quantity represents their prior belief (5/20)?
   c. What algebraic quantity indicates the likelihood of their model (5/20)?
   d. What quantity represents the evidence gathered in support or against the model (5/20)?

a $$p(M1|D1) = \frac{p(D1|M1)\,p(M1)}{p(D1|M1) + p(Di|M2)}$$

b. prior belief $p(M1)$

c. likelihood $p(D1|M1)$

d. evidence $p(D1|M1) + p(D1|M2)$

Equations are expected here

Compare theory questions chapter 5.

7. Calculate the probability of the model being correct given the fact that the data says that heart disease is present, using Bayes' rule. This is the quantity p (M1| D1). (This question is worth 10/100).

   a. Using the prior, likelihood and evidence calculate the result. This answer should be a number, and your full calculations should be clear (5/10).
   b. Restate your answer of p(M1 | D1) from question 4a. Is this the same value as your Bayes' rule calculation? Why or why not (5/10)?

a $$p(M1 \mid D1) = \frac{p(D1 \mid M1) \cdot p(M1)}{p(D1 \mid M1) + p(D1 \mid M2)}$$

$$= \frac{0.824 \cdot 0.850}{0.70 + 0.10}$$

$$= 0.876$$

b $$p(M1, D1) \mid p(D1) = p(M1 \mid D1)$$

$$= 0.876$$

Conditional probabilities and Bayes rule give the same answer as expected.

Heuristic Bayes rule would not have worked here.

If you got the wrong answer in a but knew it was wrong in b, you received credit.

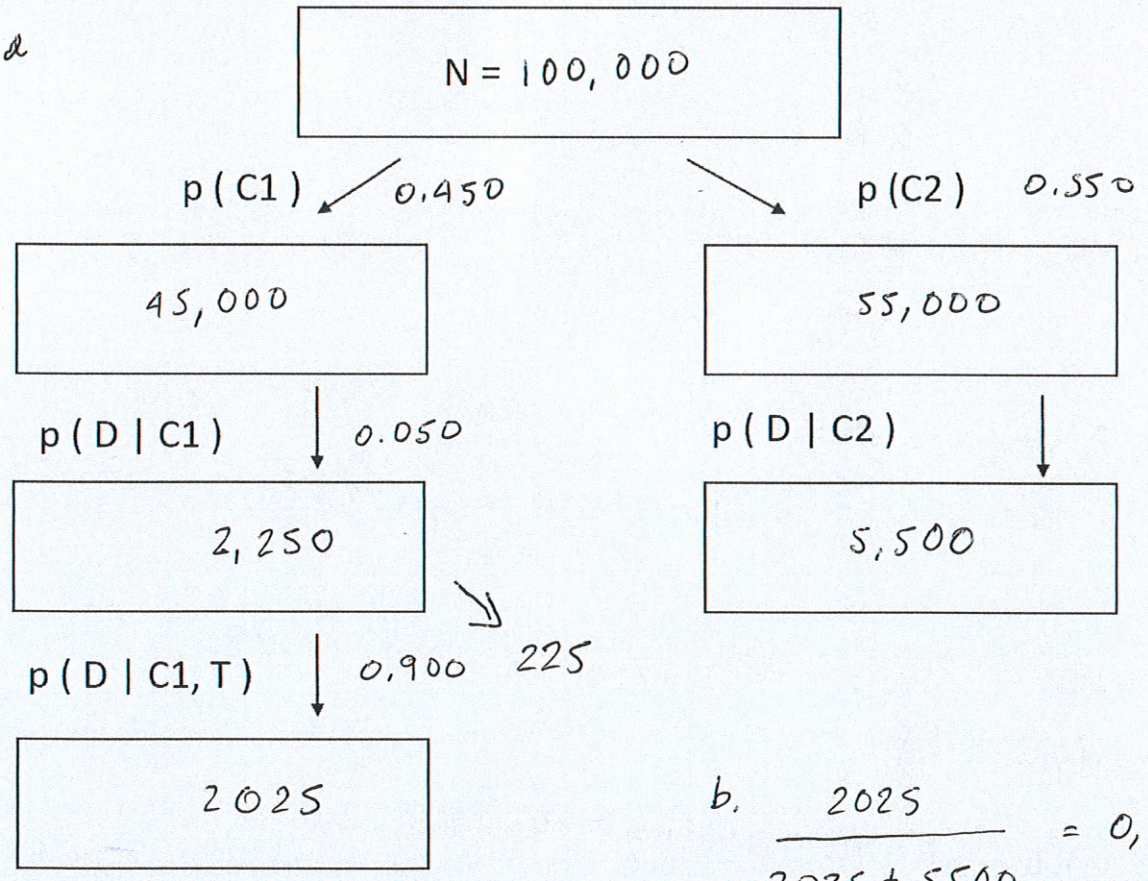
Compare prep question 8, 9

8. Suppose that there is an urban population in need of screening. The population consists of 100,000 people. Of this population 45% show up for routine checks at their GP. The other 55% do not. Of those who receive routine checks, only 5% are identified as high risk for heart disease. Of those that do not receive routine checks fully 10% are at risk of having heart disease, but go undiagnosed. If they show up to the GP, all heart disease patients identified in routine checks can be referred to a highly accurate but expensive test that will identify heart disease fully 90% of the time.

   I now describe the symbols used to represent these events in the representation below. The events of interest are "shows up to check-up" (C1), or "does not show up to check-up (C2). Another event of interest is "disease is present" (D). The final event of interest is whether the patient is "comprehensively tested" (T). (This question is worth 15/100).

   a. Use the natural frequencies or Markov representation to describe this population. Note presence of conditional probabilities such as p (D | C1, T), "the presence of the disease given check-up and exhaustive testing." Complete the attached tree, or draw your own (10/15).
   b. What percent of patients at risk are correctly diagnosed with heart disease (5/15).

a



b. $\frac{2025}{2025 + 5500} = 0,269$

This was the common answer to the question. A range of other justified answers are accepted with one point off.
Compare prep question 13, 16