Week 8

Live Discussion Session

Starts at 2.05pm

	Patients aged < 50		Patients aged 50+	
	Effective	Non-effective	Effective	Non-effective
Drug A	420	80	70	30
Drug B	85	15	150	50

Table summarizes the results from an observational study into the effectiveness of two drugs A and B for treating migraine

The 'success rate' is the percentage of effective outcomes.

Answer the following questions:

a) What was the 'success rate' for Drug A for the study participants overall? **[1 mark]** **= (420+70)/(420+80+70+30) = 81.7%**

b) What was the 'success rate' for Drug B for the study participants overall? **[1 mark]** **= (85+150)/(85+15+150+50) = 78.3%**

c) What was the 'success rate' for Drug A for the study participants aged < 50? **[1 mark]** **= 420/(420+80) = 84%**

d) What was the 'success rate' for Drug B for the study participants aged < 50? **[1 mark]** **= 85/(85+15) = 85%**

e) What was the 'success rate' for Drug A for the study participants aged 50+? **[1 mark]** **= 70%**

f) What was the 'success rate' for Drug B for the study participants aged 50+? **[1 mark]** **= 75%**

g) What can you conclude from the above results? **[2 marks]** **in each age subcategory B more effective than A, but overall A more effective**

h) Name the paradox evident in this study. **[1 mark]** **Simpson**

i) What is the main cause of the paradox in this example? **[3 marks]** **Age is a confounder. Fewer older people in study and older people more likely to take Drug B than Drug A**

j) Draw the causal model that explains the data and write down the probability tables for each node in that model. **[6 marks]**

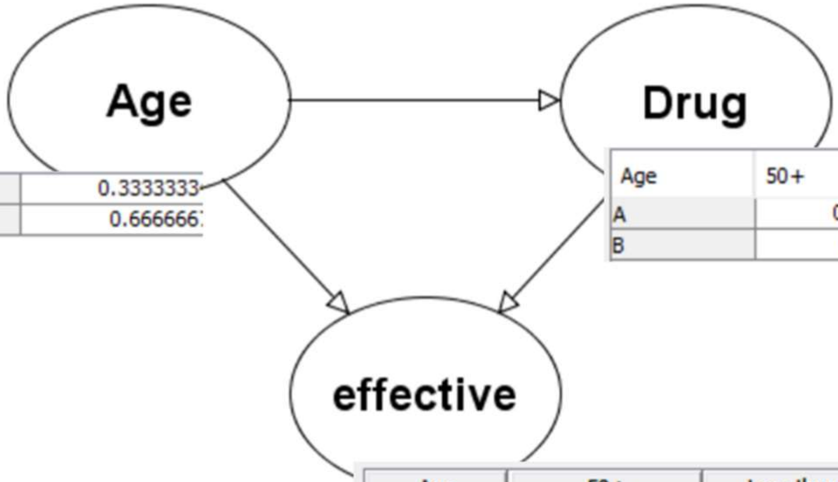
k) How would you amend the model to one that avoids the paradox? **[2 marks]**

l) By doing what you proposed in k) (or by other means) estimate the 'true' success rate for each drug for the whole population. **[4 marks]**

m) Suppose you know that a patient took Drug A and the outcome was not effective. We don't know the patient's age, but we want to answer the counterfactual question; "Would the outcome have been effective if this patient had taken Drug B instead of Drug A?". In your answer to this question provide a sketch of a causal model that supports your reasoning **[6 marks]**

	Patients aged < 50		Patients aged 50+	
	Effective	Non-effective	Effective	Non-effective
Drug A	420	80	70	30
Drug B	85	15	150	50

300
600



100	**500**
200	**100**

50+	0.3333333
Less than 50	0.666666:

Age	50+	Less than 50
A	0.33333334	0.8333333
B	0.6666667	0.16666667

Age	50+		Less than 50	
Drug	A	B	A	B
False	0.3	0.25	0.16	0.15
True	0.7	0.75	0.84	0.85

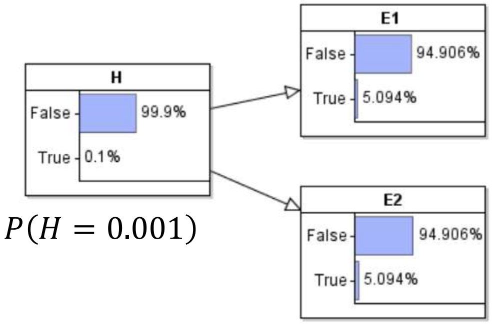
30	**50**	**80**	**15**
70	**150**	**420**	**85**

P(effective |A) = P(effective | A, 50+) x P(50+) + P(effective | A, <50) x P(<50)

A: 79.3% B: 81.7%

= 0.7 x 0.3333 + 0.84 x 0.66666 =0.793 = 79.3%

Here are our assumptions:



$P(H = 0.001)$

$P(E_1|H) = 0.99$ $\quad$ $P(E_1|not\ H) = 0.05$

$P(E_2|H) = 0.99$ $\quad$ $P(E_2|not\ H) = 0.05$

We want to calculate the probability of H if BOTH independent tests E1 and E2 are positive, i.e. calculate $P(H|(E1\ and\ E2))$

By Bayes theorem:

$$P(H|(E_1\ and\ E_2)) = \frac{P((E_1\ and\ E_2)|H) \times P(H)}{P(E_1\ and\ E_2)}$$

In the live lecture I said "Because E1 and E2 are independent we know that":

$$P((E_1\ and\ E_2\ )|H) = \ P(E_1|H) \times P(E_2|H) \qquad (1)$$

$$P(E_1\ and\ E_2) = \ P(E_1) \times P(E_2) \qquad (2)$$

But E1 and E2 are ***only*** independent once we know whether H is true or false. So, while (1) is correct, (2) is NOT correct

But, by marginalisation, we known that:

$$P(E_1 \text{ and } E_2) = P((E_1 \text{ and } E_2)|H) \times P(H) \ + \ P((E_1 \text{ and } E_2)|not\ H) \times P(not\ H)$$
$$= \ P(E_1|H) \times P(E_2|H) \times P(H) \ + \ P(E_1|not\ H) \times P(E_2|not\ H) \times P(not\ H)$$

Hence:

$$P(H|(E_1\ and\ E_2)) = \frac{P((E_1\ and\ E_2)|H) \times P(H)}{P(E_1|H) \times P(E_2|H) \times P(H) \ + \ P(E_1|not\ H) \times P(E_2|not\ H) \times P(not\ H)}$$

$$= \frac{0.99 \times 0.99 \times 0.001}{0.99 \times 0.99 \times 0.001 + 0.05 \times 0.05 \times 0.999} = 0.28183 = 28.183\%$$

Alternative method to calculate $P(H|(E1 \; and \; E2))$

$$P(H|E_1) = \frac{P(E_1|H) \times P(H)}{P(E_1|H) \times P(H) \quad + \quad P(E_1|not\ H) \times P(not\ H)} = \frac{0.99 \times 0.001}{0.99 \times 0.001 + 0.05 \times 0.999} = 0.019435 = 1.9435\%$$

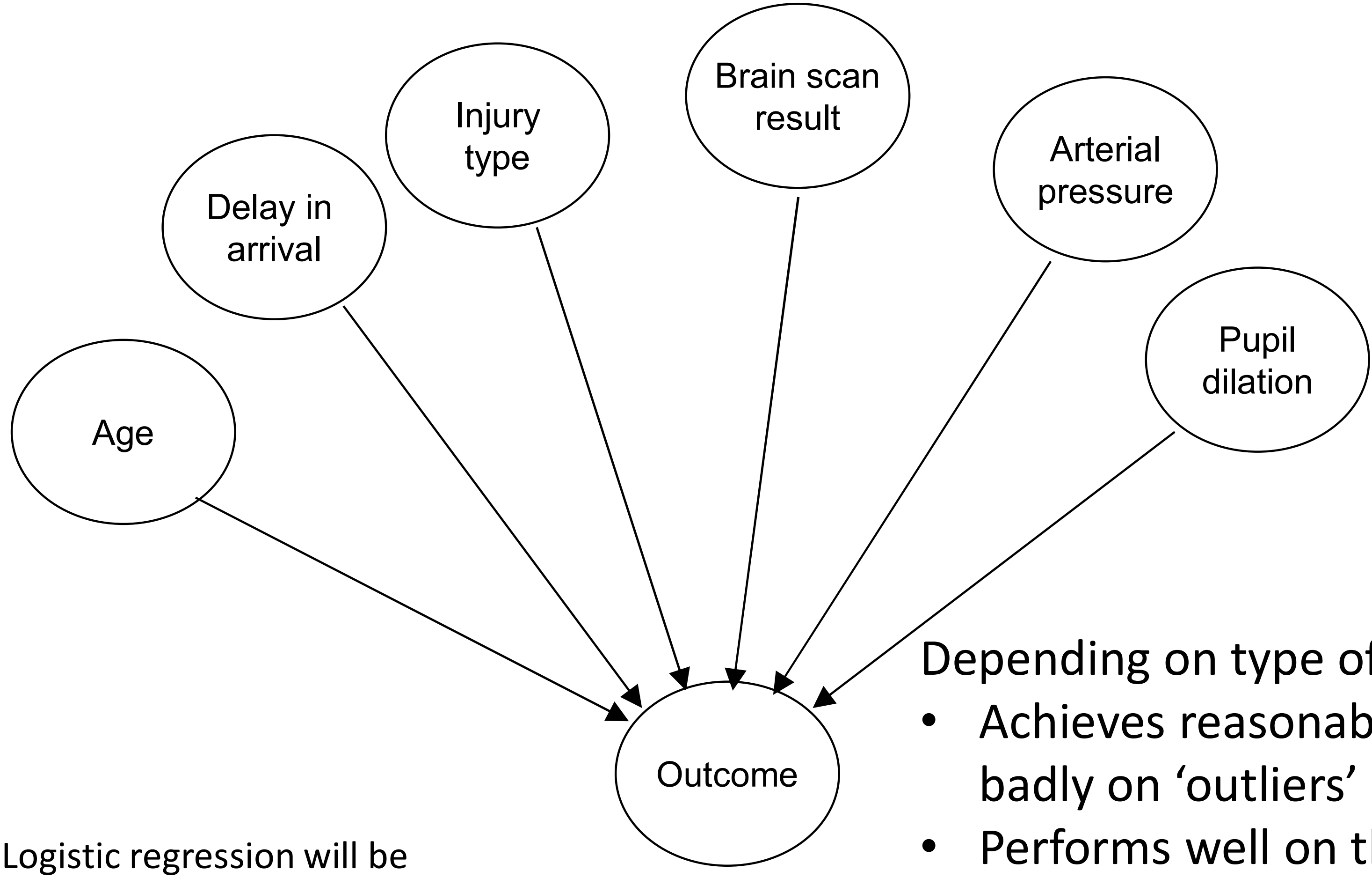
So the ‘new’ (i.e. revised probability of H) is 0.019435 instead of 0.001

Let’s call this P(H’). So this becomes the new ‘prior’ before the second test.

So:

$$P(H|(E_1 \; and \; E_2) = P(H'|E_2) = \frac{P(E_2|H') \times P(H')}{P(E_2|H') \times P(H') \quad + \quad P(E_2|not\ H') \times P(not\ H')} = \frac{0.99 \times 0.019435}{0.99 \times 0.019435 + 0.05 \times 0.980565} = 0.28183 = 28.183\%$$

Regression model* learnt purely from data ('supervised learning')

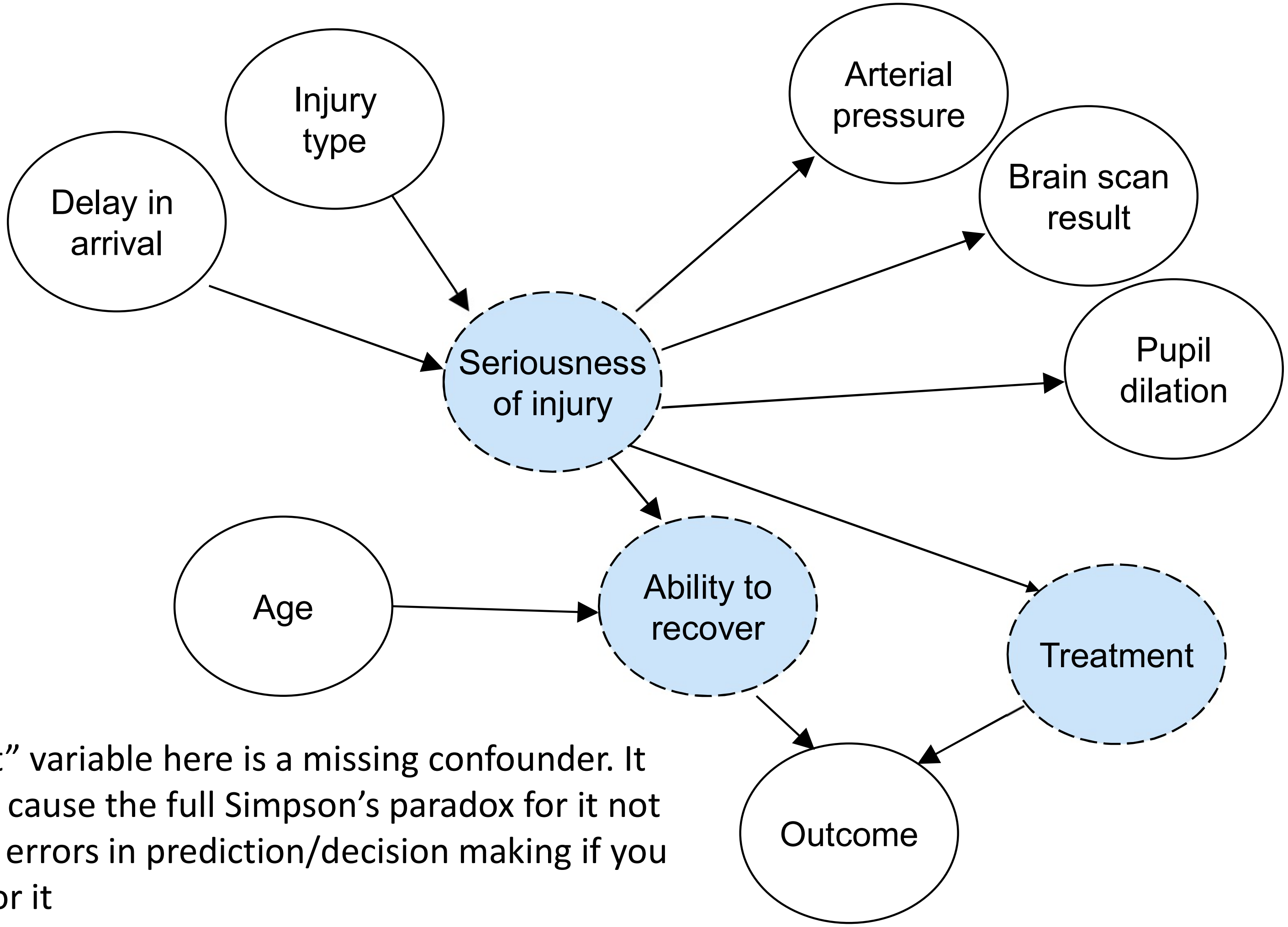


Depending on type of regression:

- Achieves reasonable accuracy but performs badly on 'outliers' (e.g fails to model 'uptick')
- Performs well on this data but is overfitted (will be poor for new data)

*Logistic regression will be covered in Lesson 8

Expert causal BN with hidden explanatory and intervention variables



More accurate (properly handles outliers), more useful for decision support

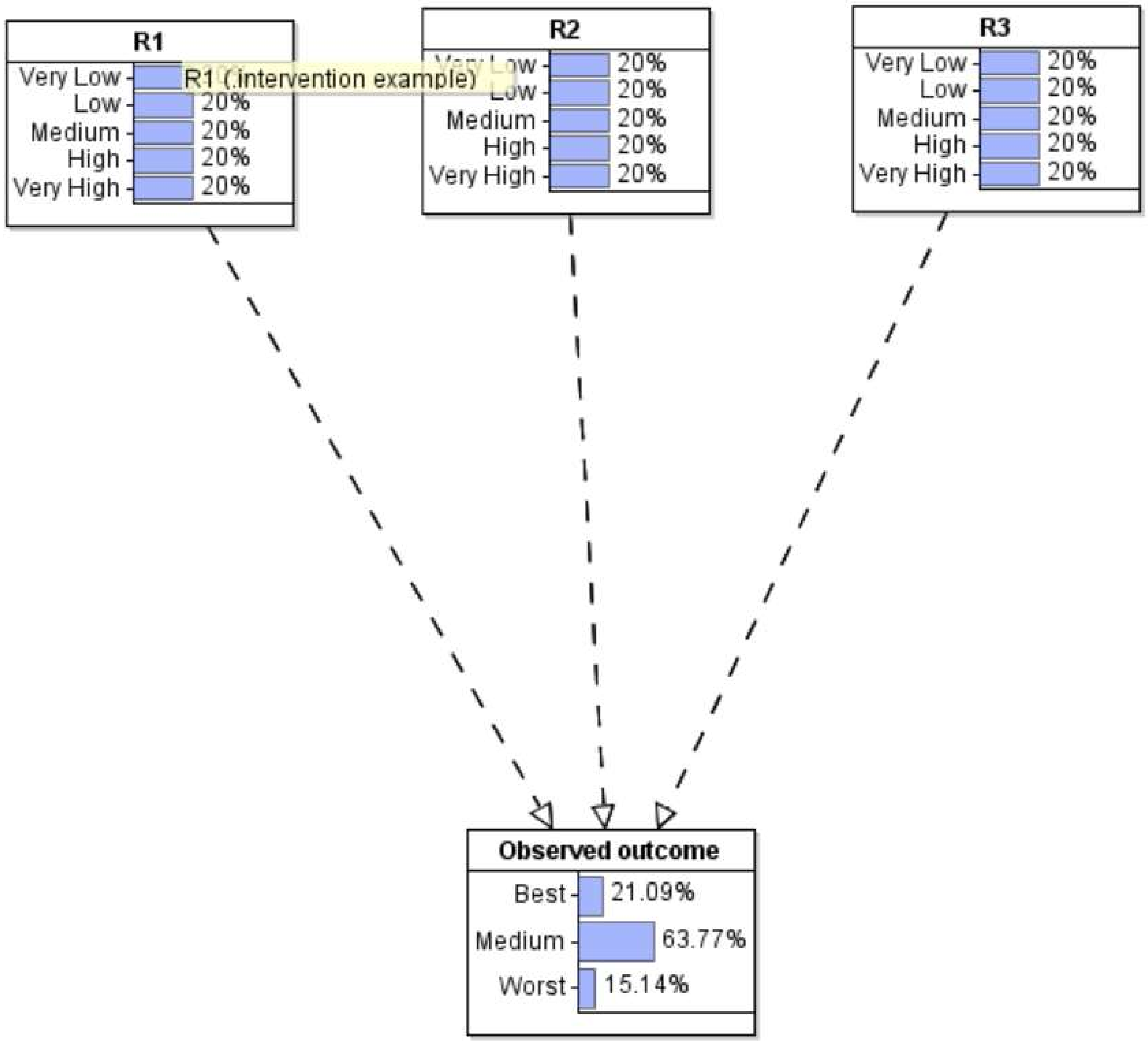
The "Treatment" variable here is a missing confounder. It doesn't have to cause the full Simpson's paradox for it not to create major errors in prediction/decision making if you fail to 'adjust' for it

WEEK 8 TASKS

Videos

Answering a counterfactual question in AgenaRisk

Life expectancy: a counterfactual example


R1
Very Low
R1 (.intervention example)
Low
20%
Medium
20%
High
20%
Very High
20%
R2
Very Low
20%
Low
20%
Medium
20%
High
20%
Very High
20%
R3
Very Low
20%
Low
20%
Medium
20%
High
20%
Very High
20%
Observed outcome
Best
21.09%
Medium
63.77%
Worst
15.14%