1. (a) Explain what a *Receiver Operating Characteristic* curve is, and how it may be used.

[2 marks]

(b) White Blood Cell count (WBC) has been proposed as a screening test for bacteremia in young children admitted to hospital with fever. To assess the accuracy of the approach, the following data were collected on 885 patients (WBC count being given in units of $1000/\text{mm}^3$).

		Bacteremia?	
		Yes	No
WBC count	0-9	2	366
	10 -14	7	293
	15 - 19	7	130
	≥ 20	10	70

(i) Taking high WBC counts to indicate bacteremia, calculate the sensitivity and specificity of the test for each cut-off level of WBC count. Plot the Receiver Operating Characteristic (ROC) curve.

[10 marks]

(ii) It is proposed to use high WBC count values to select children for further, more detailed screening tests, which are non-invasive and of reasonably low cost. Bacteremia can in 10-20% of cases lead to serious complications, including meningitis. In the light of this, which cut-off level of WBC do you consider would give the best screening test, and why?

[2 marks]

(iii) The data above are believed to be from a representative sample of young children admitted to a particular hospital with fever. Using your chosen cut-off value from (ii) above, calculate the positive predictive value of high WBC count for this population. Also compute an estimate of the prevalence of bacteremia in this population.

[2 marks]

(iv) For a different hospital, in another part of the world, prevalence of bacteremia amongst young children admitted with fever is believed to be around 10%. Using your chosen cut-off value from part (ii), calculate the positive predictive value of high WBC count when applied to this population.

[4 marks]

2. (a) Explain why *standardisation* is important when comparing mortality rates between geographical areas.

[2 marks]

(b) The table below shows the age distribution of the female population of East Berkshire in 1983, together with the numbers of female deaths in each age group in East Berkshire in 1983 and the female age-specific death rates for England and Wales during 1983.

Age group	East Berkshire female population	East Berkshire female deaths	Annual death rate for England and Wales per 1000 female population
0-1	2,400	12	9.0
1-4	9,900	5	0.4
5-14	24,100	2	0.2
15-24	28,500	8	0.3
25-34	25,900	11	0.5
35-44	24,300	27	1.2
45-54	20,400	51	3.6
55-64	19,300	163	9.6
65+	25,100	1314	53.2

(i) Draw a graph showing the annual age-specific mortality rates for females in East Berkshire and the annual age-specific mortality rates for females in England and Wales. Comment on the relationship between the two.

[6 marks]

(ii) Calculate the Standardised Mortality Ratio (SMR) for the above data and interpret your result.

[6 marks]

(iii) Calculate a 95% confidence interval for the SMR and interpret your result.

[6 marks]

3. In a study of parathyroid cancer, the times (in years) from diagnosis until death of twenty patients were recorded as follows:

   1, 1^* , 1^* , 2^* , 2^* , 3 , 5^* , 6^* , 7 , 7 , 7^* , 8 , 9^* , 10 , 10 , 11^* , 11^* , 12 , 15 , 18^*

   (* denotes a censored observation).

   (a) Calculate the Kaplan-Meier estimate of the survivor function for the above data and display your result graphically.

   [10 marks]

   (b) It is proposed to fit an Exponential distribution with mean $1/\lambda$ to the above survival time data by treating the censored observations as though they were uncensored and using the mean of all the given values to estimate $1/\lambda$. Calculate the estimated value $\widehat{\lambda}$.

   Write down the survival function $S(t)$ for an Exponential distribution with mean $1/\lambda$, and hence plot the estimated Exponential survivor function for these data on the same axes as your Kaplan-Meier estimate from part (a) above.

   [Hint: If X has an Exponential distribution with mean $1/\lambda$, then the probability density function of X is $f(x) = \lambda \exp(-\lambda x)$ for $x \geq 0$.]

   [5 marks]

   (c) Comment on the relationship between your two graphs from parts (a) and (b) above. Is the relationship as you would expect? Explain your answer.

   Would you prefer the Kaplan-Meier estimate of part (a) or the Exponential model of part (b) as a description of these data? Explain your answer.

   [5 marks]

4. (a) In a clinical trial to compare a particular treatment with no treatment with respect to ten-year mortality, it is decided to analyse the results in a Bayesian framework.

If our belief about the probability of death within ten years, θ, is represented by a Beta distribution with parameters α and β, then α and β are related to $\mathbf{E}[\theta]$ and $\mathrm{Var}[\theta]$ in the following way.

$$\begin{aligned} \alpha &= \frac{(\mathbf{E}[\theta])^2 (1 - \mathbf{E}[\theta])}{\mathrm{Var}[\theta]} - \mathbf{E}[\theta] \\ \alpha + \beta &= \frac{\mathbf{E}[\theta] (1 - \mathbf{E}[\theta])}{\mathrm{Var}[\theta]} - 1 \end{aligned}$$

For no treatment, previous evidence regarding the probability of death within ten years, θ_N, can be summarised by regarding θ_N as following a Beta distribution with mean 0.24 and standard deviation 0.02. Compute the parameters α_N, β_N of this distribution.

For the treatment under study, an expert clinician is consulted before the trial begins, and his beliefs about the probability θ_T of death within ten years on the treatment are summarised in a Beta distribution with mean 0.20 and standard deviation 0.04. Compute the parameters α_T, β_T of this distribution.

[4 marks]

(b) Express your current beliefs about the log of the odds ratio for ten-year mortality of the treatment under study compared to no treatment, and calculate a 95% credibility interval for this quantity. Hence give a 95% credibility interval for the odds ratio for ten-year mortality of the treatment compared to no treatment.

[Hint: If $\theta \sim \mathrm{Beta}(\alpha, \beta)$, then the distribution of $\log_e(\theta/(1-\theta))$ is approximately $N\left(\log_e\left(\frac{\alpha - 0.5}{\beta - 0.5}\right), \frac{1}{\alpha} + \frac{1}{\beta}\right)$.]

[6 marks]

Question 4 continued overleaf

(c) The trial was carried out and the results reported as follows. Of 20 patients randomised to the treatment, 1 had died within ten years, whereas of 20 patients randomised to receive no treatment, 4 had died. Express your beliefs following the trial about the log of the odds ratio for ten-year mortality of the treatment under study compared to no treatment, and calculate a 95% credibility interval for this quantity. Hence give a 95% posterior credibility interval for the odds ratio for ten-year mortality of the treatment compared to no treatment.

[Hint: If there are r_N deaths out of n_N patients receiving no treatment, and r_T deaths out of n_T patients receiving treatment, the posterior distribution of the log of the odds ratio is approximately Normally distributed with mean μ and variance σ^2 given by

$$\begin{aligned} \mu &= \log_e \left(\frac{(\alpha_T + r_T - 0.5)(\beta_N + n_N - r_N - 0.5)}{(\alpha_N + r_N - 0.5)(\beta_T + n_T - r_T - 0.5)} \right) \\ \sigma^2 &= \frac{1}{\alpha_N + r_N} + \frac{1}{\beta_N + n_N - r_N} + \frac{1}{\alpha_T + r_T} + \frac{1}{\beta_T + n_T - r_T} \end{aligned}$$

]

[6 marks]

(d) Comment briefly on the posterior beliefs compared to both the prior beliefs and the sample data, and on the design of the study.

[4 marks]

5. The table below shows the results of 8 randomised controlled trials comparing the drug Octreotide with a control treatment in the treatment of refractory diarrhoea.

	Octreotide		Control	
Trial	Response	No response	Response	No response
1	3	10	4	9
2	6	4	2	8
3	34	40	18	32
4	19	2	3	17
5	16	4	6	14
6	22	1	5	15
7	16	2	12	1
8	5	12	7	9

The statistic for testing for heterogeneity of treatment effects between trials is given by

$$Q = \sum_{i=1}^{k} w_i \left(\hat{y}_p - \hat{y}_i\right)^2$$

where k is the number of trials, $\hat{y}_p$ is the pooled log odds ratio across all trials, $\hat{y}_i$ is the estimated log odds ratio for trial i, and $w_i = 1/\text{Var}\left(\hat{y}_i\right)$.

[Note: $\hat{y}_i = \log_e\left[\dfrac{a_i d_i}{b_i c_i}\right]$ and $\text{Var}\left(\hat{y}_i\right) = \dfrac{1}{a_i} + \dfrac{1}{b_i} + \dfrac{1}{c_i} + \dfrac{1}{d_i}$, where

a_i denotes the number of responses to Octreotide in trial i,
b_i denotes the number of non-responses to Octreotide in trial i,
c_i denotes the number of responses to the control in trial i,
d_i denotes the number of non-responses to the control in trial i.]

(a) The pooled log odds ratio across all 8 trials is computed to be $\hat{y}_p = 0.922$. Interpret this value.

[2 marks]

(b) Test whether there is evidence of heterogeneity in the results of the trials. Interpret your results, and comment on whether it is appropriate to combine the results into a single pooled estimate of treatment effect.

[11 marks]

Question 5 continued overleaf

(c) The patients involved in trials 1, 2 and 3 were AIDS patients, whereas those in trials 4, 5, 6 and 7 were post-chemotherapy patients, and those in trial 8 were cholera patients. In the light of this, the researchers propose that if the test of part (b) above shows evidence of heterogeneity in the results of the 8 trials they will go on to test for heterogeneity within these subgroups. That is, they will conduct a test for heterogeneity on the results of trials 1, 2 and 3, and also separately conduct a test for heterogeneity on the results of trials 4, 5, 6 and 7.

Comment critically upon this suggestion.

[3 marks]

(d) What is meant by the term *publication bias*?

Describe briefly (**without** carrying out any calculations) a graphical method which could be used to investigate the possibility of publication bias in the results of the 8 trials reported above.

[4 marks]

6. (a) Explain the meaning of a *period effect* and a *carryover effect* in relation to a two-period two-treatment crossover trial.

[4 marks]

(b) In a clinical trial of a new drug for the treatment of enuresis (bed-wetting) each of 17 patients was given the drug for a period of 14 days and a placebo for a separate period of 14 days, the order of administration being chosen randomly for each patient. The table below shows the number of dry nights experienced during each treatment period.

Group 1: Drug followed by placebo			Group 2: Placebo followed by drug		
Patient	Period 1 Drug	Period 2 Placebo	Patient	Period 1 Placebo	Period 2 Drug
1	8	5	10	12	11
2	14	10	11	6	8
3	8	0	12	13	9
4	11	6	13	8	8
5	9	7	14	8	9
6	3	5	15	4	8
7	6	0	16	8	14
8	0	0	17	2	4
9	13	12			

Test the hypothesis that there is no carryover effect and interpret the result.

[12 marks]

Briefly outline the different analyses you would subsequently perform on these data if

(i) there was no evidence of a carryover effect;

(ii) there was evidence of a carryover effect.

[4 marks]

7. (a) Describe briefly the *logistic regression model*, and outline the main purposes of a logistic regression analysis.

Under what circumstances is a logistic regression analysis appropriate for survival data?

[6 marks]

(b) Explain briefly the importance of *variable selection* in regression modelling, and describe two methods by which it may be carried out.

[4 marks]

(c) For a group of 79 infants affected by haemolytic disease of the newborn, 63 survived and 16 died. For each infant the cord haemoglobin concentration and bilirubin concentration had been recorded, in the hope of using these measurements to estimate any given infant's chances of survival. A logistic regression model for survival probability was fitted to the data, with the following results.

Variable	$\hat{\beta}$	s.e.$\left(\hat{\beta}\right)$	p-value
Constant	-2.354	2.416	0.341
Haemoglobin (g/100 ml)	0.5324	0.1487	0.001
Bilirubin (mg/100 ml)	-0.4892	0.3448	0.167

Write down the fitted model. Interpret the results for the effects of the two covariates on the risk of mortality as fully as possible, and suggest what further analysis should be performed on these data.

[10 marks]