UNIVERSITY OF
CAMBRIDGE

MATHEMATICAL TRIPOS Part III

Wednesday, 6 June, 2018 9:00 am to 12:00 pm

PAPER 207

STATISTICS IN MEDICINE

Attempt no more than ***FOUR*** *questions.*

There are ***SIX*** *questions in total.*

The questions carry equal weight.

STATIONERY REQUIREMENTS

Cover sheet
Treasury Tag
Script paper

SPECIAL REQUIREMENTS

None

You may not start to read the questions printed on the subsequent pages until instructed to do so by the Invigilator.

1 Statistics in Medical Practice

(a) Consider a homogeneous continuous-time Markov process with transition probabilities $p_{rs}(t)$.

(i) Write down a formula for the expected total time $T_{rs}(t)$ spent in a state s in the period between time 0 and time t, for an individual who is in state r at time 0, in two equivalent forms: firstly, as the expectation of an integral of an indicator for an event, and secondly, in terms of the transition probabilities.

(ii) We now wish to derive a formula for the expected number of transitions $E_{rs}(t)$ that an individual, who is in state r at time 0, makes to state s before time t. Express this as the expectation of an integral of an indicator for an event.

(iii) Obtain the probability density of the event from part (ii) in terms of the transition intensities q_{rs} and/or transition probabilities of the Markov process, for example, by expressing the event as a composition of simpler events whose probabilities can be written without derivation.

(iv) Thus deduce a formula for the expected number of transitions $E_{rs}(t)$, in terms of the total time spent in each state i over the same period, $T_{ri}(t)$.

(b) Consider a chronic disease model with three states: no disease (state 1), moderate disease (state 2), severe disease (state 3), and instantaneous transitions only permitted between adjacent states. We wish to investigate how the risk of a person getting moderate disease, and the risk of progression to severe disease, is related to their age. We have a dataset, recording observations of the disease status of a set of patients, at a finite series of times. The following data are recorded for the first person.

Age (years)	
50	No disease
55	No disease
56	Severe disease

(i) Obtain an expression for this person's contribution to the likelihood of a continuous-time Markov model, in the simplest closed form, as a function of four parameters:

- the transition intensities from no disease to moderate disease, q_{12}, and from moderate to severe disease, q_{23}, for a person aged 50 years
- two parameters describing the relation between age and each of the transition intensities, which you should define explicitly.

You may use, without further simplification, an expression of the form $p_{12}(t|\lambda, \mu)$ to denote the transition probability from no disease to moderate disease over an interval of length t with constant transition intensities λ (from no to moderate disease) and μ (from moderate to severe disease), however you should substitute appropriate values of t, λ and μ.

(ii) State any simplifying assumptions being made to derive this likelihood (other than the Markov assumption). If possible, suggest why at least one of these assumptions is questionable given the data above.

(c) Suppose we expect that on average, a 50 year old without the disease will get the disease at age 60, while a 60 year old without the disease will get it at age 65. Likewise we expect that on average, a 50 year old with moderate disease will experience progression to severe disease at age 52, while a 60 year old with moderate disease will progress at age 61.

Obtain the values of the four parameters in the model from part (b) that correspond to these beliefs.

2 Statistics in Medical Practice

Consider a trial in which a new treatment is to be tested against a standard treatment. When treatment k is given to a patient (where $k = 0$ denotes the standard treatment and $k = 1$ the new treatment), a Bernoulli(p_k) outcome is observed. Because a response is desirable, p_k is the probability of a desirable outcome. Denote the total number of patients on treatment k as n_k (with $n_k \geqslant 1$) and the observed outcome from patient i on treatment k as $y_{i,k}$ for $i = 1, \ldots, n_k$. Patients will be randomised to treatments using a response-adaptive randomisation procedure. Let the allocation ratio be defined $R = \frac{n_0}{n_1}$.

(a) Show that for a fixed sample size $n = n_0 + n_1$ the allocation ratio which maximises the power of the Z-test given n and as a function of the p_k parameters (i.e. Neyman allocation) is: $R^* = \sqrt{\frac{p_0(1-p_0)}{p_1(1-p_1)}}$.
Note: the Z-test (Wald Test) is is based on a statistic $Z = \frac{\hat{p}_0 - \hat{p}_1}{\sqrt{\sigma^2_{\Delta\hat{p}}(n_0,n_1)}}$ with $\sigma^2_{\Delta\hat{p}}(n_0, n_1) = \frac{p_0(1-p_0)}{n_0} + \frac{p_1(1-p_1)}{n_1}$ where $\hat{p}_k = \dfrac{\sum_{i=1}^{n_k} y_{i,k}}{n_k}$ for $k = 0, 1$.

(b) Assume that $p_0 > p_1$. Show that if $p_0 > (1 - p_1)$ then the above derived allocation procedure assigns more patients to the inferior treatment (i.e. the new treatment).

(c) Suppose 9 patients have been already assigned to a treatment, 5 patients to the standard treatment and 4 patients to the new treatment. The observed success rates are: $\widehat{p}_0 = 3/5$ and $\widehat{p}_1 = 1/4$. Obtain expressions, in the simplest form, for the probability that patient 10 will be assigned to treatment 0 if: (i) patients are assigned using R^* as in part (a) for the allocation probabilities; (ii) the design is a randomised play the winner which started with a balanced urn with a 1:1 composition - a RPW(1,1) design; (iii) the patients are assigned optimally as derived by dynamic programming (with the goal of maximising expected successes in n patients), both arms started with a uniform prior on each p_k and patient 10 is the last patient of the trial. *Note:* there is no need to evaluate square roots.

(d) Suppose 10 patients have been already assigned to a treatment. After the 10th patient's outcome is assessed a statistical test is done at a 10% significance level and shows no significant difference between the effect of the two treatments. The trialist decides to continue to recruit another ten patients and then do a second test at a 10% significance level. Comment on the implications of this additional test on the error rate of the trial and state what part of the procedure needs to be modified to allow this to be done in a statistically robust way.

3 Statistics in Medical Practice

(a) Provide a definition of the average causal effect of binary risk factor X taking value 1 versus taking value 0 on an outcome Y using conterfactual language.

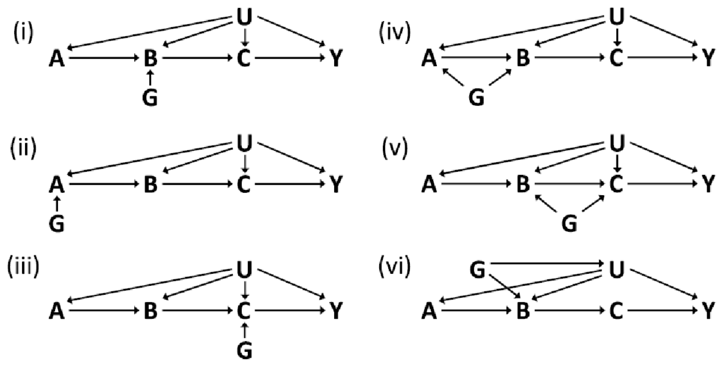
(b) What is an *instrumental variable*? What assumptions must an instrumental variable satisfy?

(c) In the causal diagrams below, in which of the six scenarios is the genetic variant G a valid instrumental variable for the risk factor B and the outcome Y (assuming that A, C and U are unmeasured)?

(d) Please read the shortened paper provided in the supplementary material over the page. What two ways do the authors of this paper try to demonstrate that elevated calcium intake is linked with increased risk of migraine?

(e) For each approach, describe three strengths and three weaknesses (potential or actual) of the approach in trying to assess the causal status of calcium as a causal risk factor for migraine.

(f) How could the authors strengthen their case in testing this causal hypothesis? Provide two suggestions, together with a justification as to how these would help strengthen causal inferences.



Supplementary material - Serum calcium and risk of migraine: a Mendellian randomization study

A migraine is a severe headache. Calcium is a nutrient that occurs in many foods and plays a vital role in the chemistry of cells. We aimed to assess the hypothesis that dietary calcium intake is a causal risk factor for migraine headaches. We first tested whether migraine headache diagnoses are associated with elevated serum calcium levels. To do this, we first obtained over 1 million de-identified health records. We observed co-occurrence between migrained headache diagnosis and hypercalcaemia (meaning excess calcium levels): odds ratio (OR) $= 1.58$ for migraine diagnosis comparing hypercalcaemia versus normal calcium levels, $P = 4.75 \times 10^{-13}$, including adjustments for age, sex, and ancestry. These data are consistent with the hypothesis that migraine and elevated calcium levels occur frequently in our patient cohort.

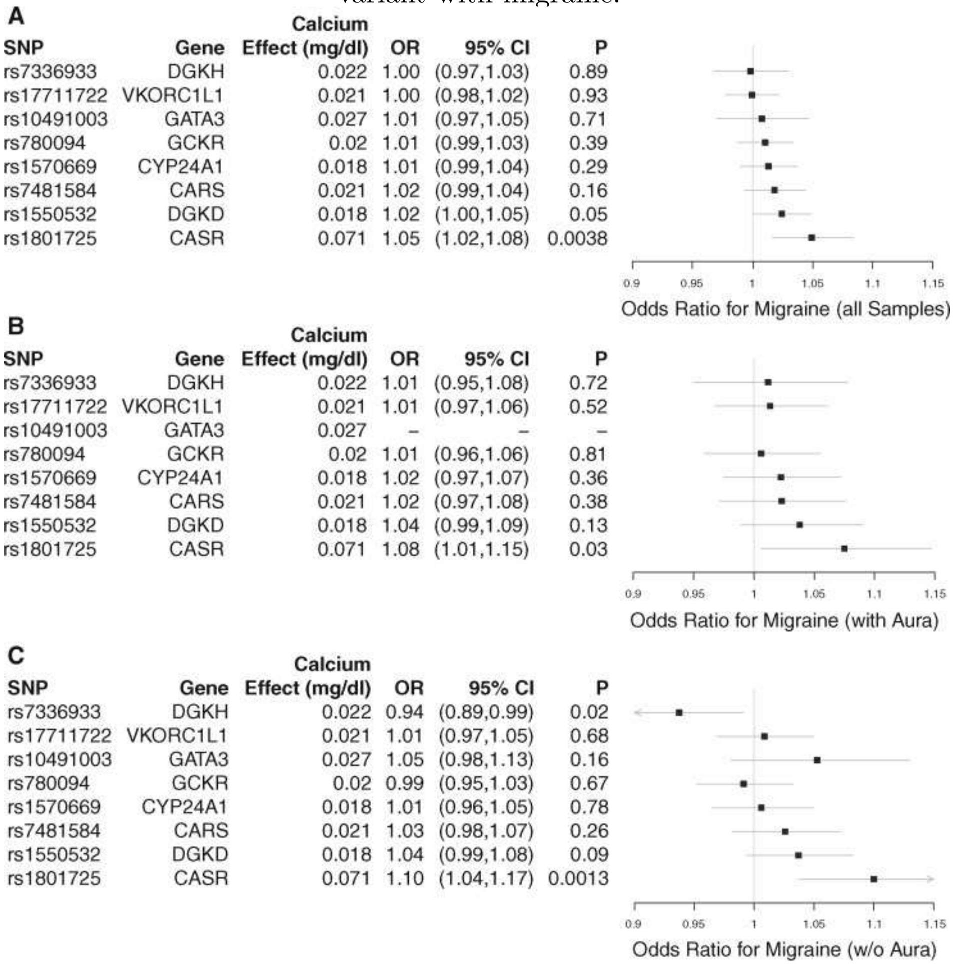
We next tested whether genetically elevated serum calcium levels are associated with increased susceptibility to migraine headache using a two-sample Mendelian randomization study design. We constructed a genetic risk score (GRS) using eight genetic variants associated with serum calcium levels (see Figure), and tested the association of this score with the outcome. The score explained 1.25% of the variance in serum calcium levels. Based on $23,285$ migraine sufferers and $95,425$ controls, we found that elevation of serum calcium levels by a hypothetical 1 mg/dL resulting from our genetic score wsa associated with an increase in risk of migraine ($OR = 1.80$, 95% CI: 1.31, 2.46, $P = 2.5 \times 10^{-4}$, see Table). We also performed sensitivity analysis methods for Mendelian randomization: the weighted mediant and MR-Egger methods.

Table: Summary estimates for genetic variants used for causal inference analysis for serum calcium for migraine traits

	Odds Ratio (95% CI)	P-value
Weighted GRS	1.80 (1.31, 2.46)	2.5×10^{-4}
Weighted-median	1.92 (1.30, 2.84)	1.6×10^{-3}
MR-Egger (Causal Effect)	1.97 (1.05, 3.69)	0.037
MR-Egger (Bias Term)	-0.003 (-0.025, 0.019)	0.74

Footnote: Odds ratios are given as change in migraine risk scaled to a unit increase in geneticall-predicted serum calcium (1 mg/dL).

Figure: Summary association data for the association of each serum calcium-associated variant with migraine.



A

SNP	Gene	Calcium Effect (mg/dl)	OR	95% CI	P
rs7336933	DGKH	0.022	1.00	(0.97,1.03)	0.89
rs17711722	VKORC1L1	0.021	1.00	(0.98,1.02)	0.93
rs10491003	GATA3	0.027	1.01	(0.97,1.05)	0.71
rs780094	GCKR	0.02	1.01	(0.99,1.03)	0.39
rs1570669	CYP24A1	0.018	1.01	(0.99,1.04)	0.29
rs7481584	CARS	0.021	1.02	(0.99,1.04)	0.16
rs1550532	DGKD	0.018	1.02	(1.00,1.05)	0.05
rs1801725	CASR	0.071	1.05	(1.02,1.08)	0.0038

B

SNP	Gene	Calcium Effect (mg/dl)	OR	95% CI	P
rs7336933	DGKH	0.022	1.01	(0.95,1.08)	0.72
rs17711722	VKORC1L1	0.021	1.01	(0.97,1.06)	0.52
rs10491003	GATA3	0.027	–	–	–
rs780094	GCKR	0.02	1.01	(0.96,1.06)	0.81
rs1570669	CYP24A1	0.018	1.02	(0.97,1.07)	0.36
rs7481584	CARS	0.021	1.02	(0.97,1.08)	0.38
rs1550532	DGKD	0.018	1.04	(0.99,1.09)	0.13
rs1801725	CASR	0.071	1.08	(1.01,1.15)	0.03

C

SNP	Gene	Calcium Effect (mg/dl)	OR	95% CI	P
rs7336933	DGKH	0.022	0.94	(0.89,0.99)	0.02
rs17711722	VKORC1L1	0.021	1.01	(0.97,1.05)	0.68
rs10491003	GATA3	0.027	1.05	(0.98,1.13)	0.16
rs780094	GCKR	0.02	0.99	(0.95,1.03)	0.67
rs1570669	CYP24A1	0.018	1.01	(0.96,1.05)	0.78
rs7481584	CARS	0.021	1.03	(0.98,1.07)	0.26
rs1550532	DGKD	0.018	1.04	(0.99,1.08)	0.09
rs1801725	CASR	0.071	1.10	(1.04,1.17)	0.0013

Calcium effect is the increase in serum calcium per additional copy of the variant allele. Odds ratio (OR) for migraine per additional copy of the variant allele. P-value (P) is for the genetic association with migraine. CI = Confidence Interval

Adapted and shortened from "Serum calcium and risk of migraine: a Mendelian randomization study" by Yin et al, HumMol Genet 2017; 26(4):820-828.

4 Analysis of Survival Data

(a) Define the *survivor function* $\mathrm{F}(t)$ for a continuous random time-to-event-variable T.

(b) A time-to-event dataset has d events in the interval $t_L < t \leqslant t_R$, with no censoring in that interval. Given that immediately after t_L there are r individuals at risk:

- (i) write down a simple estimate of the probability $\mathrm{P}\left[T > t_R | T > t_L\right]$.
- (ii) what is meant by an individual being *right-censored* at t? Why is it not possible in general to write down a simple estimate if it is known that there are c individuals right-censored in the interval?
- (iii) how would your answer to part (b)(i) differ if there are individuals right-censored at t_R but there no individuals censored in $t_L < t < t_R$?

(c) The Kaplan-Meier estimator $\hat{\mathrm{F}}(t)$ for the survivor function can be derived by considering a finite set of *potential* event times $\{a_1, \ldots, a_j, \ldots, a_g\}$ with $a_{j-1} < a_j$.

- (i) what is meant by a potential event time?
- (ii) what is the necessary condition that the set of potential event times must satisfy?
- (iii) outline the derivation of the Kaplan-Meier estimator.
- (iv) explain why the estimator does not depend on the choice of the set of potential event times, provided that condition (c)(ii) remains satisfied.

(d) Derive an alternative estimator $\tilde{\mathrm{F}}(t)$ of the survivor function by constructing a set of potential right-censoring times $\{c_0, \ldots, c_k, \ldots, c_h\}$ such that $0 = c_0 < \ldots c_{k-1} < c_k < \cdots < c_h = t$ with r_k individuals being at risk at $t = c_k$, d_0 individuals having an event at $t = 0$ and d_k individuals having an event in $c_{k-1} < t \leqslant c_k$.

Show that $\tilde{\mathrm{F}}(t) = \hat{\mathrm{F}}(t)$.

5 Analysis of Survival Data

A time-to-event dataset $\{(x_i, v_i) \colon i = 1, \ldots, n\}$ comprises n individuals: x_i being either the time of the observed event ($v_i = 1$) or the time of censoring ($v_i = 0$) for the ith individual. There are no ties in the dataset.

(a) Assume that all individuals are subject to the same hazard function. Derive the Nelson-Aalen estimator for the common integrated hazard $\mathrm{H}(t)$ in the form

$$\hat{\mathrm{H}}(t) = \sum_{i: x_i \leqslant t} \frac{v_i}{\sum_{j: x_j \geqslant x_i} 1}.$$

(b) Assume now that the ith individual is subject to hazard $\exp(\beta z_i) h_0(t)$ where $h_0(t)$ is a baseline hazard, β is a scalar parameter, and z is a scalar parameter:

(i) how would you construct a partial likelihood for β?

(ii) show that the derivative of the log partial likelihood for β is given by

$$\mathcal{S}'(\beta) = \sum_{i=1}^{n} v_i \left[z_i - \frac{\sum_{j: x_j \geqslant x_i} z_j \exp(\beta z_j)}{\sum_{j: x_j \geqslant x_i} \exp(\beta z_j)} \right].$$

(iii) interpret the expression for $\mathcal{S}'(\beta)$ in terms of the expected value of z for the individual having an event at a particular time, conditional on the history of the process to just before that time. Interpret the maximisation of $\mathcal{S}(\beta)$ to obtain an estimate $\hat{\beta}$ of β in terms of those expected values.

(iv) how would you adapt the derivation of the Nelson-Aalen estimator to provide an estimator $\hat{H}_0(t)$ of the integrated baseline hazard?

(v) the Martingale residual y_i is defined by

$$y_i = v_i - \exp(\hat{\beta} z_i) \hat{H}_0(x_i).$$

Interpret the terms on the right hand side of this equation.

(vi) outline how Martingale residuals can be used to check the functional form of an explanatory variable in a time-to-event model.

(vii) show that $\sum_{i=1}^{n} y_i = 0$. *Hint: you may find it helpful to first write out the sum for a dataset with three or four individuals.*

6 Analysis of Survival Data

Individuals are at risk of two independent events: A and B. The corresponding continuous time-to-event variables T_A, T_B have densities $\mathrm{f}_A(t)$, $\mathrm{f}_B(t)$ respectively with $\mathrm{F}_A(t) = \int_t^\infty \mathrm{f}_A(t)\mathrm{d}t$ and $\mathrm{F}_B(t) = \int_t^\infty \mathrm{f}_B(t)\mathrm{d}t$.

(i) What is the probability that $T_A < T_B$?

(ii) Derive and interpret the equation:

$$\int_0^t \mathrm{f}_A(t')\mathrm{F}_B(t')\mathrm{d}t' + \int_0^t \mathrm{F}_A(t')\mathrm{f}_B(t')\mathrm{d}t' + \mathrm{F}_A(t)\mathrm{F}_B(t) \ = 1.$$

(iii) What is the density of T_A: (1) given $T_A < T_B$ and (2) given $T_B < T_A$?

T_B is now to be interpreted as a a time-to-censoring variable with $\lim_{t\uparrow\infty} t\mathrm{F}_B(t) = 0$. The variable X is defined as $\min(T_A, T_B)$.

(iv) Find the density and expectation of X in terms of F_A and F_B.

(v) Assume that T_A has an exponential distribution with mean μ. Define a new variable U by $U = X + \mu\mathbb{I}[X = T_B]$ where $\mathbb{I}$ is the indicator function. Show that U has the same expectation as T_A.

END OF PAPER