INSTITUTE AND FACULTY OF ACTUARIES

EXAMINERS' REPORT

April 2021

Subject CS2 – Risk Modelling and Survival Analysis Core Principles Paper B

Introduction

The Examiners' Report is written by the Chief Examiner with the aim of helping candidates, both those who are sitting the examination for the first time and using past papers as a revision aid and also those who have previously failed the subject.

The Examiners are charged by Council with examining the published syllabus. The Examiners have access to the Core Reading, which is designed to interpret the syllabus, and will generally base questions around it but are not required to examine the content of Core Reading specifically or exclusively.

For numerical questions the Examiners' preferred approach to the solution is reproduced in this report; other valid approaches are given appropriate credit. For essay-style questions, particularly the open-ended questions in the later subjects, the report may contain more points than the Examiners will expect from a solution that scores full marks.

The report is written based on the legislative and regulatory context pertaining to the date that the examination was set. Candidates should take into account the possibility that circumstances may have changed if using these reports for revision.

Paul Nicholas
Chair of the Board of Examiners
July 2021

A. General comments on the *aims of this subject and how it is marked*

1. The aim of the Risk Modelling and Survival Analysis subject is to provide a grounding in mathematical and statistical modelling techniques that are of particular relevance to actuarial work, including stochastic processes and survival models.

2. Candidates are reminded of the need to include the R code, that they have used to generate their solutions, together with the main R output produced, in their answer script. Where the R code was missing from a particular question part, no marks were awarded even if the output (e.g. a graph) was included. Partial credit was awarded in the cases where the R code was included but the R output was not.

3. The marking schedule below sets out potential R code solutions for each question. Other appropriate R code solutions gained full credit unless one specific approach had been explicitly requested in the question paper.

4. In cases where the same error was carried forward to later parts of the answer, candidates were given full credit for the later parts.

5. In higher order skills questions, where comments were required, well-reasoned comments that differed from those provided in the solutions also received credit as appropriate.

B. Comments on *candidate performance in this diet of the examination.*

1. On the whole, performance was generally satisfactory. Once again, candidates typically demonstrated their ability to use R to perform analysis but did not fully demonstrate their ability to interpret the results.

2. Many candidates appeared to run out of time during the examination. In most cases this appeared to be due to candidates using very inefficient R code in their solutions, with code often running to a number of pages. This is in contrast to the examiners' solutions which may have only taken a few lines of code.

3. It is important that appropriate commentary is provided alongside the R code and R output in the answer script, where relevant, to fully demonstrate sufficient understanding. For example, in questions requiring charts, appropriate titles, axis labels and legends are necessary, and in questions requiring a specific numerical answer, this must be stated separately from the R output. Instructions to this effect were communicated to candidates at the time of the exam. Candidates are advised to take careful note of all instructions that are provided with the exam in order to maximise their performance in future CS2B examinations. The instructions applicable to this diet can be found at the beginning of the solutions contained within this document.

4. Higher order skills questions were answered very poorly. Candidates should recognise that these are generally the questions which differentiate those candidates with a good grasp and understanding of the subject.

5. Candidates are reminded that, where they are unable to answer one part of a question, the best approach is to provide a "dummy" answer and carry on with the remaining parts of the question to receive carry forward credit.

C. Pass Mark

The Pass Mark was 55.
1,422 candidates presented themselves and 480 passed.

Solutions for Subject CS2 Paper B April 2021

Please note that the following principles apply to the CS2B solutions. These principles were set out in the instructions provided to candidates along with the examination paper:

1. Candidates **MUST** include the R code used to obtain their answers in the Word document. Please note that failure to include the R code used will result in **ZERO MARKS** for that particular question.

2. Candidates **MUST** include the main R output produced from the R code in the Word document. Please note that failure to include the R output will result in full credit not being given.

3. When a question requires data to be simulated or generated in R, candidates **DO NOT** need to paste the individual values of the generated data into the Word document, unless specifically instructed to do so in the question.

4. When a question requires a particular numerical answer or conclusion, candidates **MUST** explicitly and clearly state this in the Word document, separately from, and in addition to the R output that contains the relevant numerical information. Please note that failure to include a separate answer or conclusion will result in full credit not being given.

5. Candidates should type any non-R code workings and answers into the Word document using standard keyboard typing. Candidates **DO NOT** need to use notation that requires specialised equation editing e.g. the "Equation Editor" functionality in Word.

6. Candidates **MUST** include appropriate titles, axes labels, and where relevant, legends in all graphical output that is generated in R for inclusion in the Word document. Please note that failure to include appropriate annotations will result in full credit not being given.

7. Candidates should provide relevant comments when instructed to do so in the question.

8. Your Word document **MUST NOT** contain links to any other documents.

Q1

(i)

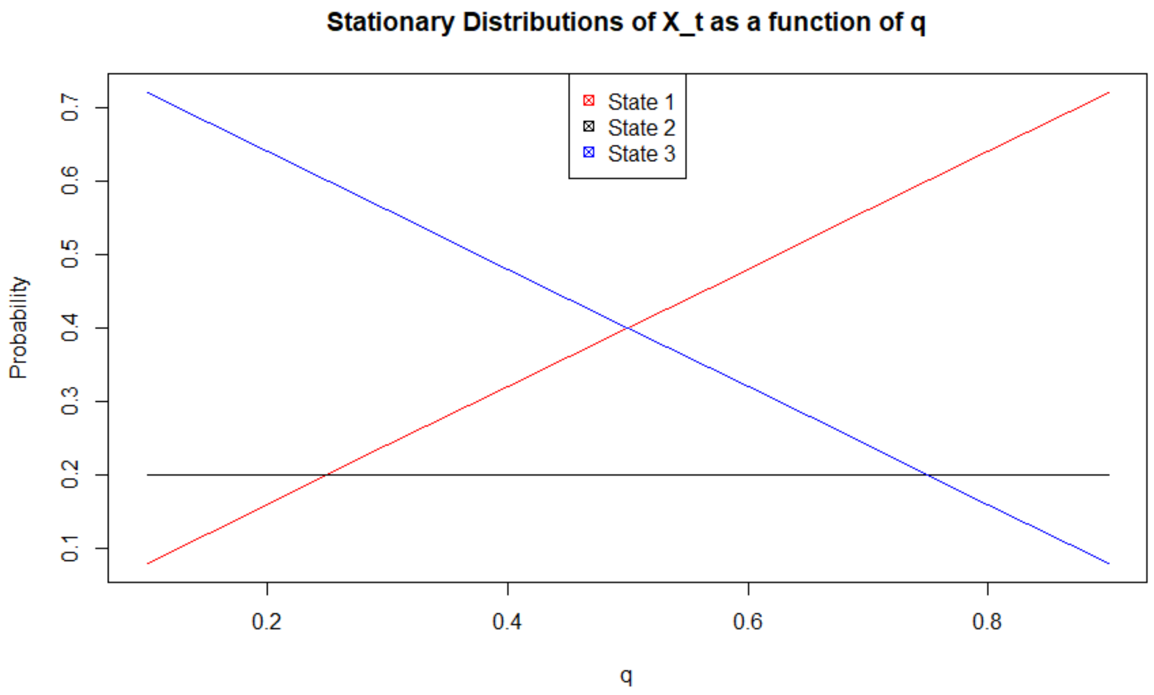
```
markovchain =
function(p,q,r){                                              [1]
P = matrix(data = c(p,1-p,0,q,0,1-q,0,r,1-r),
nrow =3, byrow = TRUE)                                        [2]
mc = new("markovchain", transitionMatrix = P)               [1½]
mc}                                                           [½]
```

(ii)

```
statdist = matrix(data = 0, nrow = 3, ncol = 9)               [1]
for(i in 1:9){                                                [1]
q = 0.1*i                                                     [1]
mc = markovchain(p = 0.75, q, r = 0.25)                       [1]
statdist[,i] = steadyStates(mc)}                              [2]
statdist                                                      [½]

     [,1] [,2] [,3] [,4] [,5] [,6] [,7] [,8] [,9]
[1,] 0.08 0.16 0.24 0.32  0.4 0.48 0.56 0.64 0.72
[2,] 0.20 0.20 0.20 0.20  0.2 0.20 0.20 0.20 0.20
[3,] 0.72 0.64 0.56 0.48  0.4 0.32 0.24 0.16 0.08
                                                              [½]
```

(iii)

```
q = seq(from = 0.1, to = 0.9, by = 0.1)                       [1]
plot(                                                         [½]
q,                                                            [½]
statdist[1,],                                                 [½]
type = "l",                                                   [½]
main = "Stationary Distributions of X_t as a
function of q",                                               [½]
ylab = "Probability",                                         [½]
col = "red")                                                  [½]

lines(q, statdist[2,], col = "black")                         [1]
lines(q, statdist[3,], col = "blue")                          [1]

legend("top",
legend = c("State 1", "State 2", "State 3"),                  [½]
col = c("red", "black", "blue"),                              [½]
pch=7)
```



[½]

[Total 20]

Part (i) was very well answered although some candidates lost marks for not creating a function, or for creating a function outputting a matrix rather than a Markov chain object. Candidates who did not create a function or who created a function outputting a matrix rather than a Markov chain object needed to perform additional work in part (ii).

Part (ii) was well answered. Candidates who were less familiar with loops in R included more extensive R code than necessary. Candidates who answered part (ii) by raising the transition matrices to a large power were required to raise the matrices to a second large power to demonstrate convergence otherwise only partial credit was awarded.

Part (iii) was poorly answered. Many candidates lost marks for missing or inappropriate titles, axis labels and/or legends. Where candidates were unfamiliar with the legend function in R, a manually applied legend was accepted ***provided*** *that it correctly represented the line colours* ***and*** *was situated next to the chart.*

Q2

```
rpareto = function(n, alpha, lambda){
rp = lambda * ((1 - runif(n))^(-1/alpha) - 1)
rp}
```

(i)

```
set.seed (123)                                                  [½]
A_vec =                                                         [½]
rpareto(n = 25000, alpha = 3, lambda = 1)                       [½]
head(A_vec, 8)                                                  [1]

[1] 0.11966335 0.67788900 0.19160373 1.04468423 1.56103061
0.01566369 0.28445342 1.10259132                                [½]
```

(ii)

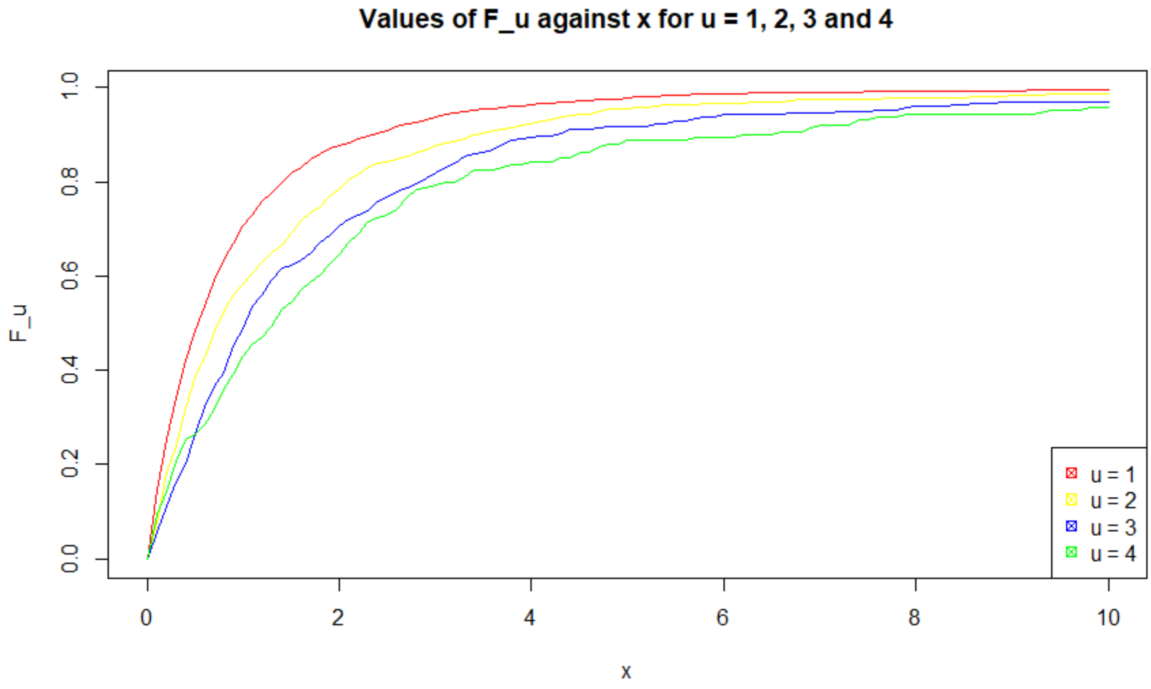
```
A_exceed_u =                                                    [½]
function(A, u){                                                 [1]
E = pmax(A - u, 0)                                             [2½]
output = E[E!=0]                                                [2]
output}                                                         [½]
head(A_exceed_u(A = A_vec, u = 1), 8)                           [1]

[1] 0.04468423 0.56103061 0.10259132 0.85069357 0.15317919
0.80118083 0.08415856 3.58825415                                [½]
```

(iii)

```
F_u =                                                           [½]
function(A_greater_than_u) {                                    [1]
y = vector(length = 101)                                        [½]
for (i in 1:101) {                                              [1]
y[i] =
length(A_greater_than_u[A_greater_than_u <= 0.1 * (i-1)]) /
length(A_greater_than_u)                                        [4]
  }
y}                                                              [½]
head(F_u(A_exceed_u(A = A_vec, u = 1)), 8)                      [1]

[1] 0.0000000 0.1463087 0.2526846 0.3439597 0.4201342
0.4838926 0.5362416 0.5895973                                   [½]
```

(iv)

```
x = seq(from = 0, to = 10, by = 0.1)                        [1]
plot(                                                       [½]
x,                                                          [½]
F_u(A_greater_than_u = A_exceed_u(A = A_vec,
u = 1)),                                                    [½]
type = "l",                                                 [½]
main = "Values of F_u against x for u = 1, 2, 3
and 4",                                                     [½]
col = "red",                                                [½]
ylab = "F_u")                                               [½]

lines(x, F_u(A_greater_than_u = A_exceed_u(A =
A_vec, u = 2)), col = "yellow")                             [2]
lines(x, F_u(A_greater_than_u = A_exceed_u(A =
A_vec, u = 3)), col = "blue")                               [1]
lines(x, F_u(A_greater_than_u = A_exceed_u(A =
A_vec, u = 4)), col = "green")                              [1]

legend("bottomright",
  legend = c("u = 1", "u = 2", "u = 3",
   "u = 4"),                                                [½]
  col = c("red", "yellow", "blue", "green"),                [½]
  pch = 7)
```



[½]

(v)
For all but the smallest values of *x*, *F_u(x)* decreases as *u* increases. [1]
This is consistent with the result that if *X* ~ *Pa*(*alpha, lambda*), then the threshold exceedance *X* – *u* | *X* > *u* is distributed as *Pa*(*alpha, lambda* + *u*). [2]
There is some irregularity caused by sampling variation due to low data volumes above the higher values of *u*. [2]
As a result of this irregularity, the curves for *u* = 3 and *u* = 4 cross over for small values of *x*. [1]

[Total 36]

Part (i) was very well answered. The most common errors were to fail to set the random number generator seed to 123 and to output the default 6 rather than 8 values using the head function.

Part (ii) was well answered. The most common error was to output all the entries of the vector E, rather than only the non-zero entries as specified in the question. Candidates are reminded to read the question carefully. Candidates unfamiliar with the pmax function included more extensive R code than necessary, referring to individual vector components.

Part (iii) was poorly answered. Few candidates coded the R function entirely accurately.

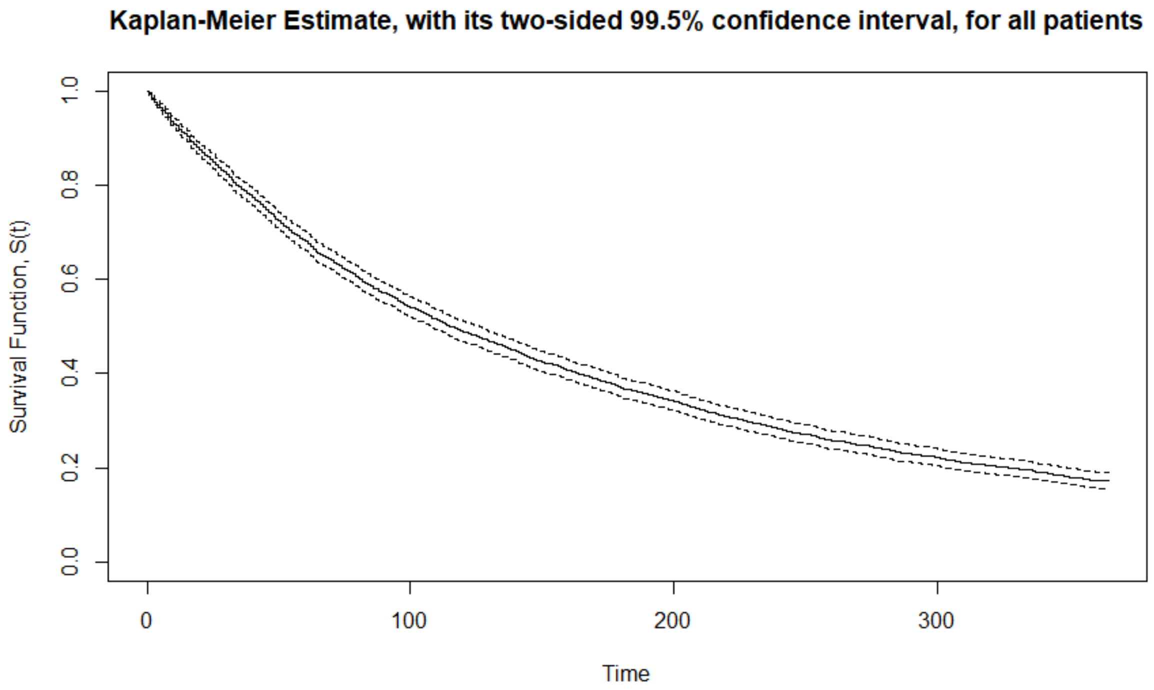
Part (iv) was surprisingly very poorly answered, despite asking for a relatively standard chart. Candidates are reminded that where they have been unable to calculate the correct data for a chart, they may still gain marks for plotting "dummy" data.

Part (v) was the least well answered question in the whole paper. Few candidates provided appropriate comments. Candidates needed to focus their comments on whether the behaviour of the curves was in line with what they would have expected based on theoretical considerations. Comments that the curves were upward-sloping did not gain credit since this is the case for any distribution function. Where incorrect values were plotted in part (iv), marks were still awarded in part (v) for any appropriate comments consistent with what was plotted.

Q3

(i)

```
mortalitydata =
read.csv(file = "CS2B_Apr_21_Qu_3_Data.csv", head = TRUE)

KMfit =
survfit(                                                      [1]
Surv(mortalitydata$Time, mortalitydata$Status) ~ 1,           [2]
conf.int = 0.995)                                             [2]

plot(                                                         [½]
KMfit,                                                        [½]
xlab = "Time",                                                [½]
ylab = "Survival Function, S(t)",                             [½]
main = "Kaplan-Meier Estimate, with its two-sided 99.5%
confidence interval, for all patients")                       [½]
```



[½]

(ii)

```
summary(KMfit, time = 365)$surv
```

<u>OR:</u>

```
KMfit$surv[365]
```

<u>OR:</u>

```
min(KMfit$surv)                                               [2]

[1] 0.1715504                                                 [½]
```

The probability that a patient survived from the beginning of the investigation to the end of the investigation is 0.172 [½]

(iii)
The probability value calculated in part (ii) is NOT suitable for assessing the effectiveness of MediCo [1]
as it is the average lifetime distribution of ALL patients in the investigation, whether they received MediCo or not [1]
To review the effectiveness of MediCo we need to compare the lifetime distribution of patients that received the drug with the lifetime distribution of patients who did not [1]

(iv)

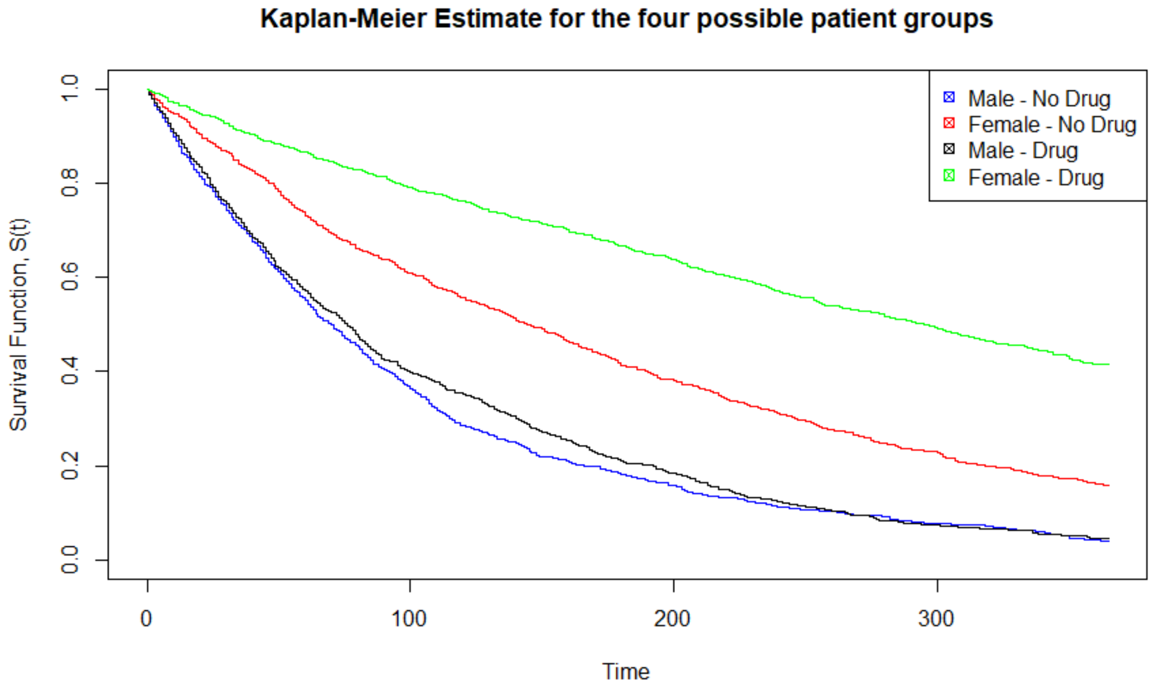
```
KMfit =
survfit(                                                  [1]
Surv(mortalitydata$Time,mortalitydata$Status)
~Drug+Gender,                                             [2½]
data = mortalitydata)                                     [1]
print(KMfit)

Call: survfit(formula = Surv(mortalitydata$Time,
mortalitydata$Status) ~
    Drug + Gender, data = mortalitydata)

                    N     events     median     0.95LCL    0.95UCL
Drug=0,    Gender=0  1100  982        70         64         76
Drug=0,    Gender=1  1100  870        145        135        158
Drug=1,    Gender=0  1100  982        76         70         81
Drug=1,    Gender=1  1100  607        295        273        317
```

<u>OR:</u>

```
KMfit =
survfit(                                                  [1]
Surv(mortalitydata$Time,mortalitydata$Status)~
mortalitydata$Drug + mortalitydata$Gender)                [3½]
print(KMfit)

Call: survfit(formula = Surv(mortalitydata$Time,
mortalitydata$Status) ~
    mortalitydata$Drug + mortalitydata$Gender)

                                              n     events      median
mortalitydata$Drug=0, mortalitydata$Gender=0    1100  982         70
mortalitydata$Drug=0, mortalitydata$Gender=1    1100  870         145
mortalitydata$Drug=1, mortalitydata$Gender=0    1100  982         76
mortalitydata$Drug=1, mortalitydata$Gender=1    1100  607         295
                                              0.95LCL     0.95UCL
mortalitydata$Drug=0, mortalitydata$Gender=0    64          76
mortalitydata$Drug=0, mortalitydata$Gender=1    135         158
mortalitydata$Drug=1, mortalitydata$Gender=0    70          81
```

```
mortalitydata$Drug=1, mortalitydata$Gender=1    273         317

     -----
```

```
plot(                                                      [½]
KMfit,                                                     [½]
xlab = "Time",                                             [½]
ylab = "Survival Function, S(t)",                          [½]
main = "Kaplan-Meier Estimate for the four possible patient
groups",                                                   [½]
col = c("blue", "red", "black", "green"),                  [½]

legend("topright",
legend = c("Male - No Drug", "Female - No Drug", "Male -
Drug", "Female - Drug"),                                   [½]
col = c("blue", "red", "black", "green"),                  [½]
pch = 7)
```



[½]

(v)
Female patients have a higher overall survival rate than male patients (with or without MediCo). [1]
MediCo seems to increase survival rates … [1]
… with the impact being much more significant on female lives than male lives. [1]

<u>EITHER:</u>
The impact of MediCo on male lives may not be statistically significant.

OR:
The impact of MediCo on male lives should be tested for statistical significance. [1]

The survival curves for males actually cross over a few times at later times, i.e. at certain points the raw estimates suggest the survival probabilities are higher for males not taking the drugs. However, this may be more reflective of a non-significant difference rather than MediCo actually reducing survival probabilities at these durations. [1]

[Marks available 5, maximum 3]

(vi)

```
coxph(                                                          [1]
Surv(mortalitydata$Time, mortalitydata$Status) ~ Drug +
Gender,                                                         [1½]
data = mortalitydata,                                           [1]
ties = "breslow")                                               [1]

Call:
coxph(formula = Surv(mortalitydata$Time, mortalitydata$Status) 
~ 
    Drug + Gender, data = mortalitydata, ties = "breslow")

          Coef      exp(coef)      se(coef)  z         p
Drug      -0.35174  0.70347        0.03439   -10.23    <2e-16
Gender    -0.93263  0.39352        0.03575   -26.09    <2e-16

Likelihood ratio test=820.3  on 2 df, p=< 2.2e-16
n= 4400, number of events= 3441                                 [½]
```

OR:

```
coxph(                                                          [1]
Surv(mortalitydata$Time, mortalitydata$Status) ~
mortalitydata$Drug + mortalitydata$Gender,                      [2½]
ties = "breslow")                                               [1]

Call:
coxph(formula = Surv(mortalitydata$Time, mortalitydata$Status) ~ 
    mortalitydata$Drug + mortalitydata$Gender, ties = "breslow")

                        coef      exp(coef)   se(coef)    z          p
mortalitydata$Drug     -0.35174  0.70347     0.03439     -10.23     <2e-16
mortalitydata$Gender   -0.93263  0.39352     0.03575     -26.09     <2e-16

Likelihood ratio test=820.3  on 2 df, p=< 2.2e-16
n= 4400, number of events= 3441                                 [½]
```

(vii)
The results suggest that MediCo reduces the mortality rate of patients by around 30% [1]
The results suggest that female lives' mortality rates are around 40% of that of males [1]
However, the graphs from part (iv) suggest that an interaction term may be required in the Cox model to fully analyse the effects of MediCo and Gender on the mortality rate, and hence the above analysis is potentially misleading [1]
The results suggest that the Gender covariate appears to have a greater effect on the mortality rate than the Drug covariate [1]
Both p-values indicate that the coefficients are unlikely to be 0 [1]

(viii)

```
coxph(                                                               [½]
Surv(mortalitydata$Time, mortalitydata$Status) ~ Drug*Gender,
                                                                     [1]
data = mortalitydata,                                                [½]
ties = "breslow")                                                    [½]

Call:
coxph(formula = Surv(mortalitydata$Time, mortalitydata$Status) ~
   Drug * Gender, data = mortalitydata, ties = "breslow")

                Coef       exp(coef)  se(coef)   z          p
Drug            -0.07143   0.93106    0.04515    -1.582     0.114
Gender          -0.63564   0.52960    0.04711    -13.494    <2e-16
Drug:Gender     -0.65874   0.51750    0.06977    -9.442     <2e-16

Likelihood ratio test=910.2  on 3 df, p=< 2.2e-16
n= 4400, number of events= 3441                                      [½]
```

<u>OR:</u>

```
coxph(                                                               [½]
Surv(mortalitydata$Time, mortalitydata$Status) ~
mortalitydata$Drug * mortalitydata$Gender,                          [1½]
ties = "breslow")                                                    [½]

Call:
coxph(formula = Surv(mortalitydata$Time, mortalitydata$Status) ~
    mortalitydata$Drug * mortalitydata$Gender, ties = "breslow")

                                             coef     exp(coef) se(coef)  z       p
mortalitydata$Drug                        -0.07143  0.93106   0.04515  -1.582  0.114
mortalitydata$Gender                      -0.63564  0.52960   0.04711  -13.494 <2e-16
mortalitydata$Drug:mortalitydata$Gender -0.65874  0.51750   0.06977  -9.442  <2e-16

Likelihood ratio test=910.2  on 3 df, p=< 2.2e-16
n= 4400, number of events= 3441                                      [½]
```

(ix)

MediCo reduces male mortality rates by 1 – 0.93106 = 6.9% [1]
which is not statistically significant [½]
and reduces female mortality rates by 1 – 0.93106 * 0.51750 = 51.8% [1]
which is statistically significant [½]
The results from the Cox analysis are consistent with the Kaplan–Meier plots in part (iv) [2]

The likelihood ratio test statistic for the model with the interaction term compared with the model without the interaction term is:
910.2 – 820.3 = 89.9

<u>OR</u>

```
L1 = m1$loglik[2]
L2 = m2$loglik[2]

-2 *(L1 - L2)
[1] 89.8897
```

(where m1 = Cox model fitted in part (vi) and m2 = Cox model fitted in part (viii))

which is highly significant under the chi-squared distribution with one degree of freedom. [2]

[Marks available 7, maximum5]

[Total 44]

Part (i) was well answered, although some candidates showed 95% confidence intervals (the default) rather than 99.5% as specified in the question. Candidates are reminded to read the question carefully.

Part (ii) was well answered, although some candidates lost marks for not including their R output or for not stating their answer separately from the output.

Part (iii) was very poorly answered with most candidates not answering the question that was asked. The question asked candidates to evaluate the **appropriateness** *of the probability value calculated in part (ii). The question was looking for an explanation that the probability value was not appropriate for assessing MediCo's effectiveness and that further information was required.*

Answers to part (iv) were generally satisfactory. However, instead of plotting a single survfit object, many candidates plotted separate survfit objects, on the same axes, for each of the four patient groups. These candidates tended to lose marks for one or more of the following reasons:

- *Defining the patient groups incorrectly, e.g. conditioning on the Gender and Drug fields separately rather than together.*
- *Producing curves that terminated too early, as not all four patient groups had death or censoring events on the last day of the investigation.*
- *Including confidence intervals, which are excluded by default when a survfit object containing multiple survival curves is plotted.*

Part (v) was very poorly answered although most candidates who produced the correct chart in part (iv) answered part (v) well.

Part (vi) was well answered although many candidates failed to specify the Breslow method for tie handling. This must be specified explicitly since the default method is the Efron method.

Part (vii) was very poorly answered. Few candidates quantified the impact of MediCo or gender on mortality rates. Very few candidates commented on the apparent need for an interaction term, which is a key conclusion from this analysis.

*Part (viii) was well answered although many candidates used Drug + Gender + Drug * Gender in place of Drug * Gender, which was unnecessary because R automatically includes the first-order terms.*

Part (ix) was very poorly answered. Many candidates commented on the statistical significance of gender in isolation, which was not relevant to this part of the question. Few candidates commented on the significance of the reduction in mortality from taking MediCo for each gender separately, or of the interaction term.

[Paper Total 100]

END OF EXAMINERS' REPORT