Final Exam
Multiple Choice Questions

1. (5 points) The data in the table below come from an early case-control study of smoking as a risk factor for lung cancer. The cases and comparable controls were sampled and queried about whether they had ever smoked at least one cigarette a day for at least a year, in which case they were defined as an ever-smoker.

	Lung Cancer	
Ever Smoker	Cases	Controls
Yes	688	650
No	21	59
Total	709	709

Based on these data, what measure would you estimate to quantify the amount of association between the study's definition of ever smoking and the risk of lung cancer?

(a) Relative risk

(b) Risk difference

(c) Odds ratio

(d) Geometric mean

(e) Median

2. (5 points) You are planning a cross-sectional study of whether mothers who began taking prenatal vitamins during the first trimester of pregnancy are less likely to give birth to a low-birthweight baby. You will sample all births taking place at a hospital over a three-month period and determine whether the mother took prenatal vitamins and if so, when she started, based on medical record review and maternal interview. You are hoping to be able to detect a risk difference associated with first-trimester prenatal vitamin use of at least -0.05. Based on preliminary data, you believe that the risk of a low-birthweight baby will be around 0.1 in women who did not begin prenatal vitamin use in the first trimester. One of your colleagues believes that this risk is actually higher, around 0.15. If you are right, will your study have more or less power to detect a risk difference of -0.05 than if your colleague is right?

   (a) More power if you are right.

   (b) More power if your colleague is right.

   (c) The power will depend only on the risk difference.

   (d) The standard error will depend on sampling variability.

   (e) The sample size will depend on which of you is right.

3. (5 points) The data below come from a preliminary report of the Physician's Health Study. Subjects are cross-classfied by whether they were randomized to regular aspirin intake or placebo and whether they experienced a fatal myocardial infarction (MI), nonfatal MI or no MI during follow-up.

	Myocardial Infarction		
	Fatal	Nonfatal	None
Placebo	18	171	10,845
Aspirin	5	99	10,933

   If you were asked to analyze these data to determine whether there was strong evidence that regular aspirin use was associated with MI-related outcomes, what hypothesis testing method would you use?

   (a) Fisher's exact test.

   (b) Cochran-Armitage test for trend.

   (c) Chi-squared test for association.

   (d) McNemar's test.

   (e) Mantel-Haenszel test.

4. (5 points) The data below come from a case-control study of alcohol as a risk factor for esophageal cancer. Subjects are cross-classified by case-control status and membership into one of four categories of daily alcohol consumption.

	Alcohol Consumption (g/day)			
	0 - 39	40 - 79	80 - 119	120+
Controls	386	280	87	22
Cases	29	75	51	45

Which of the analysis methods below would be the best for testing for an association between alcohol consumption and the risk of esophageal cancer?

(a) Generalized linear model with identity link; test of whether coefficient is zero in grouped-linear model for alcohol group.

(b) Generalized linear model with logit link; test of whether coefficient is zero in grouped-linear model for alcohol group.

(c) Generalized linear model with identity link; test of whether coefficients are zero in dummy-variable model for alcohol group.

(d) Generalized linear model with logit link; test of whether coefficients are zero in dummy-variable model for alcohol group.

(e) Weighted Kappa.

5. (5 points) Which of the analysis methods below would have been best for testing for an association between alcohol consumption and the risk of esophageal cancer if the case-control data had, instead, been as given in the table below?

	Alcohol Consumption (g/day)			
	0 - 39	40 - 79	80 - 119	120+
Controls	386	22	87	280
Cases	29	45	51	75

(a) Generalized linear model with identity link; test of whether coefficient is zero in grouped-linear model for alcohol group.

(b) Generalized linear model with logit link; test of whether coefficient is zero in grouped-linear model for alcohol group.

(c) Generalized linear model with identity link; test of whether coefficients are zero in dummy-variable model for alcohol group.

(d) Generalized linear model with logit link; test of whether coefficients are zero in dummy-variable model for alcohol group.

(e) Weighted Kappa.

6. (5 points) The data in the table below come from a matched case-control study of risk factors for myocardial infarction (MI) among Navajos. Each of 144 cases of MI was matched on age and gender to a control who was free of heart disease. Both cases and controls were asked whether they had ever been diagnosed as having diabetes. Numbers of pairs are given in the table below for each of the four case-control and diabetes-diagnosis configurations possible for a pair.

	MI cases		
Controls	Diabetes	No Diabetes	Total
Diabetes	9	16	25
No diabetes	37	82	119
Total	46	98	144

Based on these data, what is a good estimate of the age- and gender-adjusted odds ratio for the association of prior diagnosis of Diabetes with MI?

(a) $\frac{9\times82}{16\times37}$

(b) $\frac{16\times37}{9\times82}$

(c) $\frac{37}{16}$

(d) $\frac{82}{9}$

(e) Cannot adjust without more information about age and gender.

7. (5 points) Investigators analyzing the data on prostate cancer we have studied in class were interested in determining how strong the evidence was that the results of a digital rectal exam (No nodule, Unilobar nodule Left, Unilobar nodule Right, Bilobar nodule) might be associated with whether or not the tumor had penetrated the prostatic capsule.

For p = the probability that a man's prostate cancer had penetrated the prostatic capsule at diagnosis, and x giving results about what was detected by the digital-rectal exam:

$$x_2 = \begin{cases} 1 & \text{Unilobar nodule Left} \\ 0 & \text{Otherwise} \end{cases} \qquad x_3 = \begin{cases} 1 & \text{Unilobar nodule Right} \\ 0 & \text{Otherwise} \end{cases}$$

$$x_4 = \begin{cases} 1 & \text{Bilobar nodule} \\ 0 & \text{Otherwise} \end{cases}$$

the following model was fit to the data:

$$\text{logit}(p) = \beta_0 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4$$

Stata output from the fit of this model is given below.

```
. logistic penetration i.dre

Logistic regression                               Number of obs   =        380
                                                  LR chi2(3)      =      41.79
                                                  Prob > chi2     =     0.0000
Log likelihood = -235.25009                       Pseudo R2       =     0.0816

------------------------------------------------------------------------------
 penetration | Odds Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         dre |
          2  |   2.406015   .7527009     2.81   0.005     1.303189     4.44211
          3  |    4.77193   1.560659     4.78   0.000     2.513672    9.058983
          4  |   8.187135   3.163017     5.44   0.000     3.839536    17.45762
------------------------------------------------------------------------------
```

Based on this model, what can be said about the strength of evidence that the the type of nodule detected on digital-rectal exam is associated with whether the tumor has penetrated the prostatic capsule?

(a) It is strong. P = .005

(b) It is strong. P = .000

(c) It is strong. P < .001

(d) It is strong. P < .0001

(e) There is no way to tell this from the Stata output.

8. (5 points) Based on the model fit and presented in question 7, what is the odds ratio for penetration of the prostatic capsule comparing a man diagnosed with prostate cancer who had a bilobar nodule detected on digital-rectal exam to a man diagnosed with prostate cancer who had no nodule detected on digital-rectal exam?

   (a) 2.406015

   (b) 4.77193

   (c) 8.187135

   (d) 8.187135×3.163017

   (e) 5.44

9. (5 points) Based on the model fit and presented in question 7, what is the odds ratio for penetration of the prostatic capsule comparing a man diagnosed with prostate cancer who had a bilobar nodule detected on digital-rectal exam to a man diagnosed with prostate cancer who had a unilobar nodule detected on the left side on digital-rectal exam?

   (a) 2.406015

   (b) 8.187135

   (c) 8.187135×2.406015

   (d) $8.187135 / 2.406015$

   (e) 5.44

10. (5 points) Investigators were also interested in whether the tumor volume, obtained from ultrasound, was associated with whether the tumor had penetrated the prostatic capsule. Letting p = the probability that a man's prostate cancer had penetrated the prostatic capsule at diagnosis and x = tumor volume (cm^3) obtained from ultrasound, they fit the following model to the data:

$$\text{logit}(p) = \beta_0 + \beta_1 x$$

Stata output from the fit of this model is given below:

```
. logistic penetration volume

Logistic regression                                     Number of obs   =        379
                                                        LR chi2(1)      =       5.40
                                                        Prob > chi2     =     0.0201
Log likelihood = -252.92879                             Pseudo R2       =     0.0106

------------------------------------------------------------------------------
 penetration | Odds Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      volume |   .9862779    .006032    -2.26   0.024      .974526    .9981715
------------------------------------------------------------------------------
```

Based on the fit of this model, what is the estimated odds ratio associated with a one cm^3 larger tumor volume, obtained from ultrasound?

(a) 0.9862779

(b) 5.40

(c) $e^{0.9862779}$

(d) $\log 0.9862779$

(e) .006032

11. (5 points) Based on the model in Question 10, what is the estimated odds ratio associated with a 5 cm^3 larger tumor volume, obtained from ultrasound?

(a) 0.9862779

(b) 5.40^5

(c) $e^{5\times 0.9862779}$

(d) 5×0.9862779

(e) $(0.9862779)^5$

12. (5 points) A different set of investigators used the same prostate cancer data to examine the association of any nodule (left, right or bilobar) detected on digital-rectal exam with whether the tumor had penetrated the prostatic capsule, and how this association depends on age.

For p = the probability that a man's prostate cancer had penetrated the prostatic capsule at diagnosis,

$$x_N = \begin{cases} 1 & \text{Nodule (left, right or bilobar) detected} \\ 0 & \text{Otherwise} \end{cases}, \quad \text{and}$$

x_A = age in years,

they fit the following model to the data:

$$\text{logit}(p) = \beta_0 + \beta_1 x_N + \beta_2 x_A + \beta_3 x_N x_A$$

The Stata output below gives the results of this model fit:

```
. logistic penetration i.nodule##c.age

Logistic regression                                     Number of obs   =        380
                                                        LR chi2(3)      =      26.81
                                                        Prob > chi2     =     0.0000
Log likelihood = -242.73783                             Pseudo R2       =     0.0523

------------------------------------------------------------------------------
 penetration | Odds Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
    1.nodule |   18.68739   56.31742     0.97   0.331      .050853    6867.207
         age |   1.018341   .0415079     0.45   0.656      .940152    1.103033
             |
nodule#c.age |
          1  |   .9766058   .0436343    -0.53   0.596     .8947217    1.065984
------------------------------------------------------------------------------
```

Based on the Stata output from this model, what is the strength of evidence that a nodule detected on digital rectal exam is associated with the tumor having penetrated the prostatic capsule?

(a) Not strong. P = .656

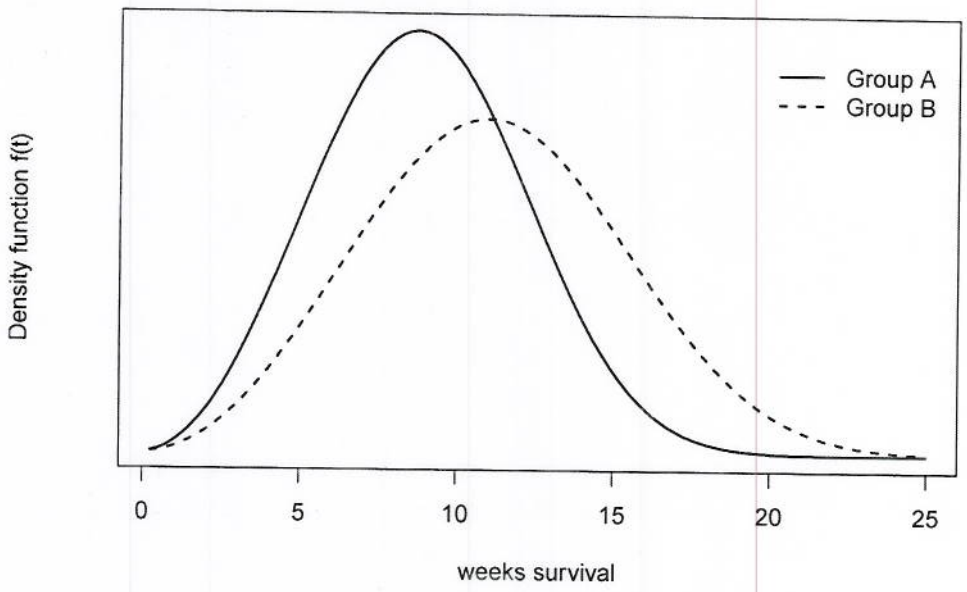
(b) Not strong. P = .331

(c) Not strong. P = .596

(d) Strong. P < .0001

(e) Cannot tell from the information given on the output.

13. (5 points) The plot below shows the density functions governing the distribution of survival times after diagnosis, in weeks, in two groups of patients diagnosed with a terminal illness.



Based on this plot, which group do you think is likely to have the higher hazard of death at most times?

(a) Group A

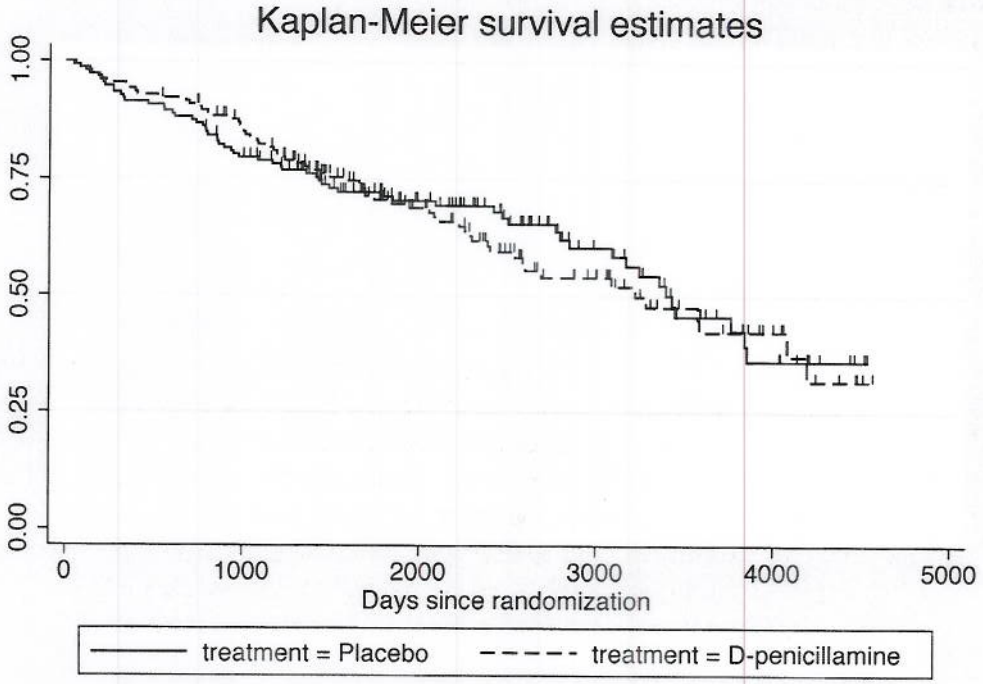
(b) Group B

(c) Neither one; the hazard functions obviously cross each other.

(d) It is impossible to even guess, given the information in this plot.

(e) The hazards are higher earlier in time.

14. (5 points) The next three questions concern data from a double-blind randomized clinical trial of D-penicillamine vs. Placebo in patients at the Mayo Clinic who had primary biliary cirrhosis (PBC). This trial was conducted in the 1970s and 1980s, before liver transplantation became an effective treatment for PBC. From 1974-1984, PBC Patients meeting eligibility criteria who presented at the Mayo clinic were randomly assigned to one of the two trial arms and followed until they died or until the time of trial data analysis in 1986. The Figure below presents Kaplan-Meier survival functions, by treatment group, for the event of death from any cause.



Based on this plot, which treatment has the higher estimated median survival time, D-penicillamine or Placebo?

(a) D-penicillamine

(b) Placebo

(c) It is impossible to tell from the plot.

(d) The median survival times are identical.

(e) The median survival times are not reached in these data.

15. (5 points) The abbreviated Stata output below shows results from fitting a Cox regression model to the data on death from any cause in the PBC clinical trial described in the previous question.

Letting $x_T = \begin{cases} 1 & \text{D-penicillamine} \\ 0 & \text{Placebo} \end{cases}$, the following model was fit:

$$\lambda(t) = \lambda_0(t)e^{\beta x_T},$$

```
. stcox treatment

...

Cox regression -- Breslow method for ties

No. of subjects =          312                     Number of obs    =       312
No. of failures =          125
Time at risk    =       625985

                                                   LR chi2(1)       =      0.10
Log likelihood  =   -639.92903                     Prob > chi2      =    0.7498

------------------------------------------------------------------------------
          _t | Haz. Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
   treatment |   1.058787   .1896977     0.32   0.750     .7452519     1.50423
------------------------------------------------------------------------------
```

Pretend that this proportional hazards model fits the data well. Based on the hazard ratio estimate in this model, which treatment group is estimated to have better survival?

(a) D-penicillamine

(b) Placebo

(c) The estimated survival distributions are identical.

(d) You cannot tell from the output which group has better estimated survival.

(e) There is no survival curve estimate provided in this output.

16. (5 points) The abbreviated Stata output below shows results from fitting a second Cox regression model to the data on death from any cause in the PBC clinical trial. Letting x_T be defined as in the previous question, and letting x_B = serum bilirubin level at randomization (mg/dl), the following model was fit:

$$\lambda(t) = \lambda_0(t)e^{\beta_T x_T + \beta_B x_B}$$

```
. stcox treatment bilirubin

...

Cox regression -- Breslow method for ties

No. of subjects =          312                     Number of obs    =       312
No. of failures =          125
Time at risk    =       625985

                                                   LR chi2(2)       =     85.79
Log likelihood  =   -597.08411                     Prob > chi2      =    0.0000

------------------------------------------------------------------------------
          _t | Haz. Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
   treatment |   1.222253   .2241719     1.09   0.274     .8531852    1.750971
   bilirubin |   1.163459   .0154566    11.40   0.000     1.133556    1.194151
------------------------------------------------------------------------------
```

Pretend that this proportional hazards model fits the data well. Why is the hazard ratio associated with treatment so much higher in this model than it is in the model in question 15 (that does not include serum bilirubin level)?

(a) Serum bilirubin level at randomization confounds the association between treatment and survival.

(b) Serum bilirubin level at randomization is strongly associated with subsequent survival.

(c) Serum bilirubin level is in the causal pathway.

(d) There was random variation in the data.

(e) Treatment improves survival.

17. (5 points) This question refers to the model first examined in question 12. For your convenience, variable coding and Stata output are repeated below.

For p = the probability that a man's prostate cancer had penetrated the prostatic capsule at diagnosis,

$$x_N = \begin{cases} 1 & \text{Nodule (left, right or bilobar) detected} \\ 0 & \text{Otherwise} \end{cases}, \quad \text{and}$$

x_A = age in years,

they fit the following model to the data:

$$\text{logit}(p) = \beta_0 + \beta_1 x_N + \beta_2 x_A + \beta_3 x_N x_A$$

The Stata output below gives the results of this model fit:

```
. logistic penetration i.nodule##c.age

Logistic regression                                     Number of obs   =        380
                                                        LR chi2(3)      =      26.81
                                                        Prob > chi2     =     0.0000
Log likelihood = -242.73783                             Pseudo R2       =     0.0523

------------------------------------------------------------------------------
 penetration | Odds Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
    1.nodule |   18.68739   56.31742     0.97   0.331      .050853    6867.207
         age |   1.018341   .0415079     0.45   0.656      .940152    1.103033
             |
nodule#c.age |
           1 |   .9766058   .0436343    -0.53   0.596     .8947217    1.065984
------------------------------------------------------------------------------
```

Based on this model, what is the estimated OR (for the tumor having penetrated the prostatic capsule) associated with finding a nodule on the digital-rectal exam, among 50-year-olds?

18. (5 points) Data in the table below cross-classify subjects by how two different pathologists classified the same slides of cervical tissue according to the presence and, if present, the extent of cervical cell abnormalities.

	Pathologist B				
Pathologist A	Negative	Atypical Squamous Hyperplasia	Carcinoma *in situ*	Squamous or Invasive Carcinoma	Total
Negative	22	2	2	0	26
Atypical Squamous Hyperplasia	5	7	14	0	26
Carcinoma *in situ*	0	2	36	0	38
Squamous or Invasive Carcinoma	0	1	17	10	28
Total	27	12	69	10	118

What measure would you use to characterize the amount of agreement between the two pathologists? Please be as explicit as possible.

19. (5 points) The Stata output below gives results of fitting a logistic model to data from an unmatched case-control study of alcohol as a risk factor for oral cancer.

Letting p = the probability a subject is a case,

x_{AGE} = (age in years), and

$$x_{ALC} = \begin{cases} 1 & \text{alcohol consumption at least 80 g/day} \\ 0 & \text{otherwise} \end{cases},$$

the following model fit the data well:

$$\text{logit}(p) = \beta_0 + \beta_1 x_{AGE} + \beta_2 x_{ALC}$$

```
. logistic case age alcohol

Logistic regression                               Number of obs   =        975
                                                  LR chi2(2)      =     179.68
                                                  Prob > chi2     =     0.0000
Log likelihood =  -404.9061                       Pseudo R2       =     0.1816

------------------------------------------------------------------------------
        case | Odds Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         age |   1.063514   .0077538     8.45   0.000     1.048425     1.07882
     alcohol |   5.929853   1.109397     9.51   0.000      4.10956    8.556429
------------------------------------------------------------------------------
```

Based on the results presented in this output, write a one- or two-sentence presentation of the study results, such as would be appropriate for the results page of the scientific article. (If you write more than two sentences, I will only read the first two.)

20. (5 points) The abbreviated Stata output from a Cox regression model below gives results comparing radiation and chemotherapy to radiation only in a randomized clinical trial of treatments for cancer of the oropharynx. In the model below, a number of prognostic factors were included by using true statification on unique combinations of values of the following variables: tumor grade (well, moderately or poorly differentiated), tumor site (faucial arch, tonsillar fossa, posterior pillar, pharyngeal tongue, posterior wall), tumor size (2 cm or less largest diameter, 2 - 4 cm largest diameter with minimal infiltration in depth, more than 4 cm diameter, massive invasive tumor) and tumor stage (no clinical evidence of node metastases, single positive node 3 cm or less in diameter–not fixed, single positive node more than 3 cm in diameter–not fixed, multiple positive nodes or fixed positive nodes).

Letting x_S hold a unique value for each combination of the stratification variables listed above, and letting

$$x_T = \begin{cases} 1 & \text{Chemotherapy and radiation} \\ 0 & \text{Radiation only} \end{cases}$$

the following model for the hazard function for death from any cause fit the data well:

$$\lambda(t) = \lambda_{0x_S} e^{\beta x_T}$$

```
. stcox trtgrp, strata(grade site size stage)

        failure _d:  status
  analysis time _t:  time
...
Stratified Cox regr. -- no ties

No. of subjects =          194                     Number of obs    =       194
No. of failures =          141
Time at risk    =       108589
                                                   LR chi2(1)       =      1.19
Log likelihood  =   -127.55324                     Prob > chi2      =    0.2745

------------------------------------------------------------------------------
          _t | Haz. Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      trtgrp |   1.305469   .3199754     1.09   0.277     .8074843    2.110568
------------------------------------------------------------------------------
                                          Stratified by grade site size stage
```

Based on this Stata output, on the next page, write a one- to two-sentence description of the study results such as would be appropriate for the results section of the scientific paper reporting the clinical trial. (If you write more than two sentences, I will only read the first two.)

20. (5 points) Please write your answer to question 20 here.