EPID 716: Epidemiologic Data Analysis
Syllabus
Spring, 2016

Instructor:
Christy Avery
Christy_avery@unc.edu
Office (off-campus): 137 E. Franklin St., Suite 306, Chapel Hill, NC 27514
Office hours: 9:00 – 9:50 AM, Wednesdays, Hooker atrium

Lecture: All students, Wednesday, 8:00-8:50, 1305 McGavran-Greenberg

Labs and Teaching Assistants:
Sections 602-604 meet in 307 or 329 Health Sciences Library. Your TA will post a schedule on SAKAI. Section 601 meets in 304 Carrington Hall.

1:	Suzanne Landi	(601: Wednesday, 1:25-3:15)
2:	Marie Stoner	(602: Wednesday, 2:30-4:20)
3:	Kathleen McClain	(603: Thursday, 10:00-11:50)
4:	Kate Rucinski	(604: Thursday, 2:00-3:50)

Laboratory Website:
All assignments, data etc. will be posted in the "Lab" folder on the EPID 716 Sakai site.

Outline:

A. Overview and General Information

Objectives:
EPID 716 consists of two related analytic exercises: a cohort study and a case-control study. The primary objectives of the laboratory component are:

1. To develop an understanding of basic data analytic procedures
2. To improve data analytic skills
3. To develop skills for epidemiologic abstract and manuscript preparation
4. To reinforce epidemiologic concepts presented to date

Cohort Study Analyses (Independent/Small Group Assignments)
During weekly laboratory sessions, you will conduct a cohort study of preterm birth using 2012 North Carolina Live Birth Certificate data. Preterm birth is a major cause of perinatal mortality and morbidity and is a major economic burden to families and society. Previous studies have demonstrated that the risk of preterm birth is lower

among women who receive early prenatal care than among women who do not receive early care, but funding for North Carolina public health programs to improve and expand prenatal care is limited. Therefore, the primary aims of the cohort study analysis are:

1. To estimate the magnitude of the association between early prenatal care and preterm birth in North Carolina in 2012
2. To determine whether the association between prenatal care and preterm birth differs among major racial/ethnic populations that might be targeted for state funded prenatal care programs.

Preterm birth is defined as a live birth during the 17-week interval beginning with the 21st week of gestation and ending when 37 weeks of gestation are completed. The exposure of primary interest is prenatal care during the first 20 weeks of gestation. Race (of the mother and child, classified using US Census categories), mother's Hispanic ethnicity, smoking during pregnancy, mother's age, and child's sex will be evaluated as potential effect measure modifiers or confounders of the relation between prenatal care and preterm birth. To conduct the study you will complete six written assignments that require you to:

- prepare data from the 2012 North Carolina live birth certificate data files for analysis;
- describe the study population;
- estimate risks, rates, and measures of effect;
- conduct dose-response analyses;
- evaluate and adjust for confounding;
- study effect measure modification;
- interpret results;
- and examine the sensitivity of your results to analytic assumptions.

Case-Control Study (Independent Final Project)

The final lab project is based on an analysis of a second North Carolina birth certificate data set. This individual project will require you to apply and extend methods and skills developed for the cohort study assignments. Results of the case-control project are presented in an abstract with supplemental tables and a figure. Instructions for the final case-control project will be available in early April. You are strongly encouraged to begin the analysis early.

Background Information

Links to the following references are available on the EPID 716 Sakai site (see the Lab folder under Course Documents). Please read these before your first lab session for background information relevant to the lab project.

- Berkowitz & Papiernik. Epidemiology of preterm birth. *Epidemiologic Reviews* 1993;15:414-443.
- Buescher PA, et al. The quality of the new birth certificate data: A validation study in North Carolina. *AJPH* 1993;83:1163-5.

- Savitz D A, Hertz-Picciotto I, Poole C, Olshan AF. Epidemiologic measures of the course and outcome of pregnancy. *Epidemiol Rev, 24:* 91-101, 2002 (focus on the introduction, the section on preterm birth, and the discussion.)
- The North Carolina Live Birth Certificate

Acceptable outside resources

For the orientation assignment: You may consult books, coding manuals, online materials, and coursework. However, you may **not** consult coursework from previous EPID 716 course material, the laboratory component of EPID 715 (prior to 2013), EPID 268 course material, the EPID 700 assignments that reference NC live birth certificate data, or any previous SAS competency/orientation exams. Please note that use or reference to prohibited coursework (including coursework completed in a previous year if retaking EPID 716) will be immediately reported as a violation of the honor code.

For cohort assignments 1-6 and the case-control project: You may consult books, coding manuals, online materials and prior coursework from courses other than the EPID 268 course material, the laboratory component of EPID 715 (prior to 2013) and EPID 716 materials (2013 forward). Please note that use or reference to prohibited coursework (including coursework completed in a previous year if retaking EPID 716) will be reported as a violation of the honor code.

If you have questions regarding what is considered an appropriate source, please ask your TA or Christy.

Attendance

We do not take attendance during lecture or lab, but expect you to attend your assigned laboratory section. Your TA will be available during the lab to answer questions about assignments, programming, etc., and you will benefit from working with other students. In addition you are responsible for any new topics, corrections, and instructions that are presented or announced during lecture or lab.

Statistical Support

Basic statistical programming support will be provided in SAS. Please discuss any concerns about your level of statistical programming expertise with your TA or Christy. It is important that your skill level is high enough to facilitate composing and documenting the SAS code yourself. Handouts posted on the class website will include programming suggestions for most lab assignments. ***However, you should add documentation and examine the code carefully to ensure that you understand all of it.*** Please feel free to use and explore other ways to complete analyses and other resources, either online or print.

Additional Programming Resources

Please review the additional SAS code/tips/tricks provided on the Sakai site. We also will note the first instance when these additional resources are useful on the first page of each assignment.

Errata
To ensure that typos and points of clarification are communicated to all students, we will maintain errata under the "Background, syllabus, and errata" folder on Sakai. A notification will be emailed via Sakai every time the errata are updated.

Notebook
Notebooks are critical tools for analysis. You should maintain a lab notebook that includes your notes, programs, results, and lab assignments. You should also document decisions made during the analysis, including changes in coding, etc. Printed program files and output from SAS will be major components of your lab notebook.

Group Work
We strongly encourage you to work in groups on your cohort study analysis and assignments. A group is defined as no more than three students who are enrolled in the same lab section. You may choose your own group (i.e. groups are not assigned.) As a group, you may turn in a single written assignment, for which each group member will receive the same grade or you may work with your group to perform analyses and interpret results, but submit your laboratory assignments individually. Either way, you must be sure you understand **all** aspects of the work and are capable of completing all of the analyses and assignments on your own. This is especially important because the final lab project (the case-control analysis) **must** be done on an individual basis.

Grading: Cohort Assignments
- All cohort study assignments will be graded and returned by your TA.
- If assignments are submitted as a group, all members of the group will receive the same grade.
- "Unauthorized assistance" for cohort assignments consists of computer programming, completed assignments, and answer keys from previous years.
- In total, the cohort assignments are worth 75% of your final 716 grade.

Grading: Case-control Project
- The final case-control project is worth 25% of your 716 grade. It will be graded by your TAs and Christy.
- Case-control projects must be completed individually. "Unauthorized assistance" for the case-control project consists of computer programming, completed assignments or answer keys from previous years, and consultation pertaining to the specifics of the assignment with anyone other than a TA or Christy.

Policy for Late Assignments
Each cohort assignment is due at the **beginning** of your lab session, with the exception of the case-control assignment, which should be placed in the EPID 716 box by 3:00 PM on Wednesday, April 27th. Your TA may grant an extension for an individual assignment if you request the extension at least one business day in advance, have an approved reason for the delay, and agree with your TA on the date that the assignment

will be submitted. Christy and the TAs will decide whether "proof" documenting your absence (e.g. obituary, doctor's note etc.) is required. Any dishonesty in requesting an extension will be immediately reported as an honor code violation.

Cohort assignments

If you do not submit your assignments on time (unless an extension is granted), points will be deducted from the grade you would have received had you submitted the assignment when due. Specifically, 5% of the point total per day will be deducted for each day an assignment is overdue. Assignments that are more than 7 days overdue will receive a grade of 0.

> Orientation assignment: Cohort assignment 1 is not considered submitted (i.e. is subject to standard late policies) until the student earns 80% or better on the orientation assignment. This policy does not apply to students who passed EPID 700 in 2015, or the summer 2015 SAS competency exam. If students are submitting cohort assignment 1 as a group, we will apply the late policy on an individual basis.

Case-control project

The final case-control project should be placed in the EPID 716 box by 3:00 PM on Wednesday, April 27nd. Five percent per day will be deducted from your final grade if the project is submitted up to five days late. Case-control projects submitted after 3:00 PM on Monday, May 2nd will receive a grade of 0.

Grading

The six cohort projects account for 75% of the class grade. Each of the six cohort projects are weighed according to the number of available points. The case-control assignment accounts for the remaining 25%. The grade scale is:

<60.0	F
60.0 – 74.9	L
75.0 – 89.9	P
≥ 90.0	H

Style Points

95% of your grade for all assignments is based on content. The remaining 5% is assigned based on style. Factors considered when assigning style points include the formatting of graphs and tables, rounding rules, and legibility in hand calculations. Please note that style points only refer to the way you present results. Text exceeding word limits, errors resulting from the use of rounded values in computations, and other such oversights/errors are graded as content, not style.

Instructor availability

Please plan ahead if you need to contact an instructor outside the regularly scheduled lecture (all instructors), office hours (Avery), and lab (Landi, McClain, Rucinski, Stoner) times. We will do our best to reply to emails within 24 hours on weekdays. You may email instructors on weekends or during vacation, but we cannot promise to respond

until the following Monday. In general, the following tips will help you obtain the most helpful and prompt response.

- Direct your email to only one instructor. We strongly prefer that students contact his/her TA or Christy.
- Use emails to ask brief questions and request clarifications on assignments. Referring to specific labs, page numbers, etc. in the body of your email also is helpful.
- Do your best to condense questions into one email.
- Carbon copy (cc) group members on emails if the question pertains to a shared assignment.
- If you have difficulty condensing your questions into a handful of items, consider scheduling a meeting with an instructor or attending Christy's office hours.
- Difficulty understanding concepts is often best addressed in person.

Honor Code

Every assignment, both group and individual efforts, **must** be submitted with a signed honor code stating that: "**On my honor, I have neither given nor received unauthorized aid on this assignment**". If the assignment is a group effort, each group member must submit a separate pledge. As an alternative, each student may sign the same document listing the pledge. All assignments are not considered complete until they are accompanied by the honor code signed by all members of the group (when applicable). Please note that under the Instrument of Student Judicial Governance, the usual sanction for a first time academic offense is suspension for one academic semester and a failing grade. No exceptions will be made. (More information on the honor code is available at: http://honor.unc.edu/)

EPID 716 Syllabus, Spring 2016

	Topic		
Dates	**Lecture**	**Lab**	**Assignments Due***
1/13– 1/14	Lab introduction, orientation to data set and cohort analysis, introduction of TAs	Orientation to lab data†	----------
1/20 – 1/21	Data description/writing for an epidemiologic audience	1. Data Description (Group work begins)	Orientation assignment†
1/27 – 1/28	General and generalized linear models	2. Crude Analyses: Categorical Variables	Assign. 1: Data Description
2/3 – 2/4	General and generalized linear models	2. Crude Analyses: Categorical Variables	----------
2/10 – 2/11	Analysis of continuous variables	3. Crude Analyses: Continuous Variables	Assign. 2: Categorical Variables
2/17 – 2/18	Analysis of continuous variables	3. Crude Analyses: Continuous Variables	----------
2/24 – 2/25	Confounding, midterm evaluations	4. Confounding	Assign. 3: Continuous Variables
3/2 – 3/3	Confounding, DAG analysis	4. Confounding	----------
3/9 – 3/10	Effect measure modification, review midterm evaluations	5. Effect measure modification	Assign. 4: Confounding
3/16-3/17	***----------Spring Break (No lecture or lab)----------***		
3/23 – 3/24	Effect measure modification	5. Effect measure modification	----------
3/30 – 3/31	Sensitivity analysis	6. Sensitivity Analyses	Assign. 5: Effect measure modification
4/6 – 4/7	Sensitivity analysis	6. Sensitivity Analyses	----------
4/13 – 4/14	Case-control studies	Case-control project	Assign. 6: Sensitivity analysis
4/20 – 4/21	Case-control studies, course evaluation	Case-control project	----------
4/27– 4/28	***No lab or lecture***		Case-control project due in class mailbox by 3 PM, 4/27

*Assignments are due at the beginning of your scheduled lab section, with the exception of the case-control project (see syllabus). †Students who completed either EPID 700 in 2015 or the August, 2015 version of the SAS competency exam are exempt.

C. General Instructions for Laboratory Assignments

There are six written assignments for the cohort analysis. Please read all instructions before beginning each section of analysis. SAS suggestions will be included in the instructions or provided in separate handouts. If you need additional help with programming, ask a fellow student or your TA. If you are a good programmer, please help mentor others in your group.

Please observe the following general rules:

1. Always keep a backup copy of the original (unaltered) data set.
2. Always save your programs and output.
3. Never hand in raw data, logs, or output files that come directly from your software package. You may cut and paste output tables into another document, but only if the data are formatted, labeled, and annotated so that they are easy to understand and interpret. It is not acceptable to turn in the raw computer logs with hand written or poorly annotated comments.
4. Do not refer to variables by their "software" names (e.g., "**mage**" or "**mage2**") when formally presenting results or describing analyses. Such names are convenient for instructions and lab notebooks, but they will be meaningless to outside readers. Over time they will become meaningless to you as well!
5. Unless instructions specifically state that you must answer a question in a manner that would be "suitable for publication" you may use bulleted lists, incomplete sentences, etc., but your answer must be clear to the reader. Refer to results papers published in epidemiologic journals (such as *Epidemiology* and the *American Journal of Epidemiology*) for examples of text that would be "suitable for publication" when you are specifically instructed to provide this type of response.
6. Points will be deducted when answers exceed maximum length requirements specified. Be clear and concise. If you can answer in fewer words than the maximum allowed, please do.

Appropriate Use of Examples

The lab syllabus and instructions for each assignment are an important and appropriate source of information and examples for written assignments, in addition to tables and figures provided as examples of appropriate formatting and labeling. We encourage you to use these examples and information sources as a starting point for your work; however, it is not acceptable to cut and paste or otherwise plagiarize text from the syllabus, lab instructions or other sources for written assignments, or to copy table footnotes, titles, or other features into new tables that you are supposed to create on your own. *If you have any questions about what is acceptable, please check with your TA or Christy.*

Hand Calculations

Some assignments require hand calculations. The purpose is to help you understand the calculations and procedures used in data analysis. You will need a calculator with *ln* and e^x functions and the capacity to store at least one value. You may use a spreadsheet program, such as Excel, to perform "hand calculations", as long as you

create your own spreadsheet for this purpose. Use of Excel spreadsheets from prior semesters is not allowed.

Rounding

- To simplify grading we may sometimes require that you use specific rounding rules, including those based on significant figures; otherwise you can use the rounding rule(s) of your choice.
- If no rounding rule is given, choose your own, but be consistent.
- Avoid using rounded values in computations and retain as many decimal places as possible during intermediate steps (i.e., more than you will use when reporting your final result).
- Round up for 5's
    - Example: 0.525 = 0.53 (rounded)

Significant figures review

Example: rounding to two significant digits

Not rounded	Rounded
0.008822	0.0088
0.08822	0.088
0.8822	0.88
8.822	8.8
88.22	88

For more practice using significant digits, see
http://science.widener.edu/svb/tutorial/sigfigures.html

General Instructions for Tables

"Fill in the blank" tables are sometimes provided for written assignments. When you are asked to generate your own table(s), follow the formatting instructions below. These instructions were adapted from the *American Journal of Epidemiology*:
http://www.oxfordjournals.org/our_journals/aje/for_authors/general.html

Table Format and Style Requirements

- Tables should be concise and self-explanatory. Table titles should give details on the place of study, the time of the study, and the study population (if applicable). The designation TABLE 1 should be typed flush left, followed by a period and the title. In the title, capitalize only the first word and proper nouns. (In text, use a lowercase beginning letter for "table" and "figure.")
- Use a single top rule, a single rule below the headings, and a single bottom rule. Do not use rules within the table body. Column headings should be clearly delineated, with straddle rules over pertinent columns to indicate subcategories. Whenever possible, data in vertical columns should have the same unit of measurement.
- In the table body, leave blank spaces for no entry; avoid using dashes.

- For decimal fractions less than 1.0, use a zero in the whole-number position (e.g., 0.01).
- Place confidence intervals in a separate column from point estimates, without parentheses.
- Footnotes should include information on the derivation or definitions of variables, methods used to generate data, and any other information needed to ensure that the table can be interpreted independently, without referring to the manuscript text. For footnotes, use the following symbols in this order:
  *, †, ‡, §, #, **, ††, ‡‡, etc.

General Instructions for Figures

- Figures should be self-explanatory. The figure legend should note the time and place of the study, a brief description of the study design, population and method(s) used to generate data shown (as appropriate), and any other information needed to interpret the figure.
- Define all figure abbreviations in the legend.
- Letters, numbers, decimal points, and symbols should be clear and large enough to read easily.
- Always use a logarithmic scale when plotting ratio measures (e.g., relative risks, odds ratios).

D. Description of the Cohort Study Data

Description of the Cohort Study Data

The cohort study population consists of 64,616 eligible North Carolina live births for which the birth and the entire 17-week risk period for preterm birth (the 21st week of gestation through the 37th week of gestation) occurred in 2012.

Source data

Data are from the 2012 North Carolina Live Births file. We have provided you with a SAS version of this dataset. For additional references or interests outside this class (i.e. we have provided an EPID 716 version of these data for you on SAKAI, do not use anything other than this dataset), you may download these data with documentation from the Odum Institute for Research in Social Sciences: go to Data Archive Services, Vital Statistics). All data included in the file are taken directly from North Carolina live birth certificates, which are filed by all NC hospitals and birthing centers. Copies of the NC birth certificate and documentation for the complete NC Live Birth Certificate dataset are posted in the Cohort Study Dataset folder.

Eligibility criteria

The equivalent of a closed cohort was constructed from birth certificate data for 2012. In a closed cohort, everyone must be at risk of the outcome at entry into the cohort, and all members of the cohort must remain at risk until they experience the outcome or complete the entire risk period for the outcome. The risk period for preterm birth begins after 20 weeks of gestation have been completed (i.e., on the first day of the 21st week of gestation) and ends when a preterm birth occurs, or when 37 weeks of gestation

have been completed. Live births that occur on or after the first day of the 38th week of gestation are term births.

North Carolina live birth certificates were filed for 122,513 births in 2012. Of these, 64,616 met the cohort study eligibility requirements described below.

1. At least 20 weeks of gestation completed before birth.
   - Fetuses born before 20 weeks of gestation were completed were never at risk of preterm birth (i.e., they were never at risk of live birth during the 21st through the 37th week of gestation).
     - Gestational age is defined by the live birth data file variable "Completed Weeks of Gestation (calculated)" (**weeks**), which is calculated from the "date last normal menses began" and the date of birth*. This gestational age variable does not take into account any other information (e.g., from ultrasound examinations) that might be used to refine estimates of gestational age.
     - The variable **weeks** indicates the number of weeks of gestation completed at birth, so that births with **weeks** = 20 were preterm births during the 21st week of gestation, births with **weeks** = 36 were preterm births during the 37th week of gestation, and births with **weeks** = 37 were term births during the 38th week of gestation.
   - 39 births occurred before 20 weeks of gestation were completed (**weeks** <20).
   - 137 births were missing data for gestational age at birth.

   *By convention, days of gestation are usually counted from the first day of the last normal menstrual period (LMP) rather than the exact date of conception (which is rarely known).

2. Fewer than 20 weeks gestation completed before January 1, 2012.
   - This eligibility criterion ensures that all births during the17-week risk period for preterm would have been observed in 2012 (i.e., would have been included in the 2012 NC Live Birth Certificate dataset).
     - To meet this criterion, the LMP date had to occur after August 13, 2011 To confirm this we first determined the week of year 2012 in which each birth occurred ('**weeknum**' 1-53, such that **weeknum** = 1 for all births occurring Jan. 1-7, and **weeknum** = 53 for births Dec. 30-31). If **weeks** – **weeknum** was >19, the birth was ineligible.
   - 42,180 births completed 20 or more weeks of gestation before Jan. 1, 2012.

3. Start date of gestation at least 45 weeks prior to Jan. 1, 2013 to ensure that all eligible births (preterm or term) were observed during 2012.
   - The oldest gestational age in the data set is 45 weeks. Pregnancies that began less than 45 weeks prior to Jan. 1 2013 might have occurred in 2013; therefore the latest LMP date for an eligible birth was Feb. 25, 2012 (the last day of the 7th week of 2012). Using the available data, we defined births as eligible if **weeknum** – **weeks** was <=8.
   - 12,643 births had an approximate start date < 45 weeks prior to Jan. 1, 2013.

4. Singleton births only.
    - Multiple gestation is a strong risk factor for preterm birth. We excluded multiple gestation births from the data set (identified based on the live birth data file variable '**plural**') since etiologic mechanisms leading to preterm birth in multiple gestation pregnancies probably differ from those leading to the majority of preterm births.
    - 4,207 births were not singleton pregnancies.
    - 0 births were missing data on multiple gestation status.

5. No congenital anomalies recorded on the birth certificate.
    - Congenital malformations are associated with preterm birth. We excluded births with congenital anomalies since etiologic mechanisms may differ from other preterm births.
    - Congenital anomalies listed on the birth certificate and used as exclusion criteria were: anencephaly, meningomyelocele/spina bifida, cyanotic congenital heart disease, congenital diaphragmatic hernia, omphalocele, gastroschisis, limb reduction defect (excluding congenital amputation and dwarfing syndromes), cleft lip with or without cleft palate, cleft palate alone, Down syndrome, suspected chromosomal disorder, hypospadias. Births without data for congenital anomalies were also excluded.
    - 809 births had congenital anomalies.
    - 244 births were missing data regarding congenital anomalies.

Note:

- The exclusion criteria are applied sequentially.
- Characteristics of the final study population (e.g., numbers of eligible births) may change after the data are "cleaned" (i.e., examined for consistency, out of range values, missing data, etc.).

Limitations

1) We assumed that the following events (competing risks) were rare enough to be negligible:

- Fetal deaths after 20 weeks gestational age.
- Women moving into or out of North Carolina after completing the 20th week of pregnancy.

2) All data are from birth certificates. Accuracy cannot be confirmed, and some data are incomplete. Criteria for recording birth certificate data may vary somewhat. For example, race/ethnicity should be self-reported, but may be classified by the person completing the birth certificate in some cases. Information regarding gestational age is recorded in one-week intervals and is based on a variety of error-prone methods used to estimate the length of gestation.

- Terms and descriptors used in the EPID 716 data sets and instructions are consistent with terms used in the data collection instrument (the birth certificate) and source data files.

E. Coding Manual for the Cohort Study Data Sets

Variable name	*Description*	*Coding*
weeks	Completed weeks of gestation (calculated)	20-45, 99 99 = unknown
prenatal	Month of pregnancy when prenatal care began	1-9, 99 1-10 = month prenatal care began 88 = No prenatal care 99 = unknown
race	Race of mother/child	1-4 1 = White 2 = African American 3 = American Indian or Alaska Native 4 = Other
hispmom*	Hispanic origin of mother	N = No Y = Yes U = Unknown
cigdur*	Cigarette smoked during pregnancy	N = No Y = Yes U = Unknown
sex	Gender of child	1 = male 2 = female 9 = unknown
mage	Age of mother (years)	12-50 99=unknown
idnum	Identification number	1-64621
weeknum**	Year 2012 week of birth	1-53 1 = Jan. 1-7, 2012 2 = Jan. 8-14, 2012 …. 53 = Dec. 30-31, 2012

Variables were taken directly from the source data (North Carolina Live Births, 2012) except for **idnum** and **weeknum**. * **hispmom** and **cigdur** are coded as string (i.e. non-numeric) variables. ** **weeknum** was derived from the date of birth, which is not included in the cohort dataset. Week 53 (Dec. 30-31) includes two days.