

Pharmacometric Analyses in Clinical Trials using R

FAES BioTech84

April 25-28, 2017

Course Outline (4 days)

- **Day 1**

- General Introduction
- R Installation
- Intro to Pharmacokinetics
- Using R for PK (NCA)
- Clinical Trial Design

- **Day 2**

- Statistics for Clinical Trial Analyses
- Exposure/Response modeling I
 - Linear models
- Exposure/Response modeling II
 - Binary data, OR, RR, Logistic Regression

- **Day 3**

- Exposure/Response modeling III
 - Ordinal (polytomous) data
 - Proportional odds models
 - Count data
 - Poisson regression models
 - GEE
 - GLMM

- **Day 4**

- Exposure/Response modeling IV
 - Survival data
 - Cox models
- Summary, Conclusions

Day 1

9:00 - 9:15am:

- General introduction

9:15 – 10:15am:

- R Installation

10:15-10:30am:

- Break

10:30-12:00pm:

- Introduction to Pharmacokinetics

12:00 – 1:00pm:

- Lunch break

1:00 – 3:00pm:

- Using R to perform NCA
- Correlating PK data w/
Demographics in R

3:00 - 3:15pm:

- Break

3:15 – 4:30pm:

- Clinical trial design

General Introduction – Drug Development

- Early drug development – a simplistic overview:
 1. Identification of a **disease** to treat
 2. Identification of a **target** that affects disease
 3. Design of **compounds** to affect target

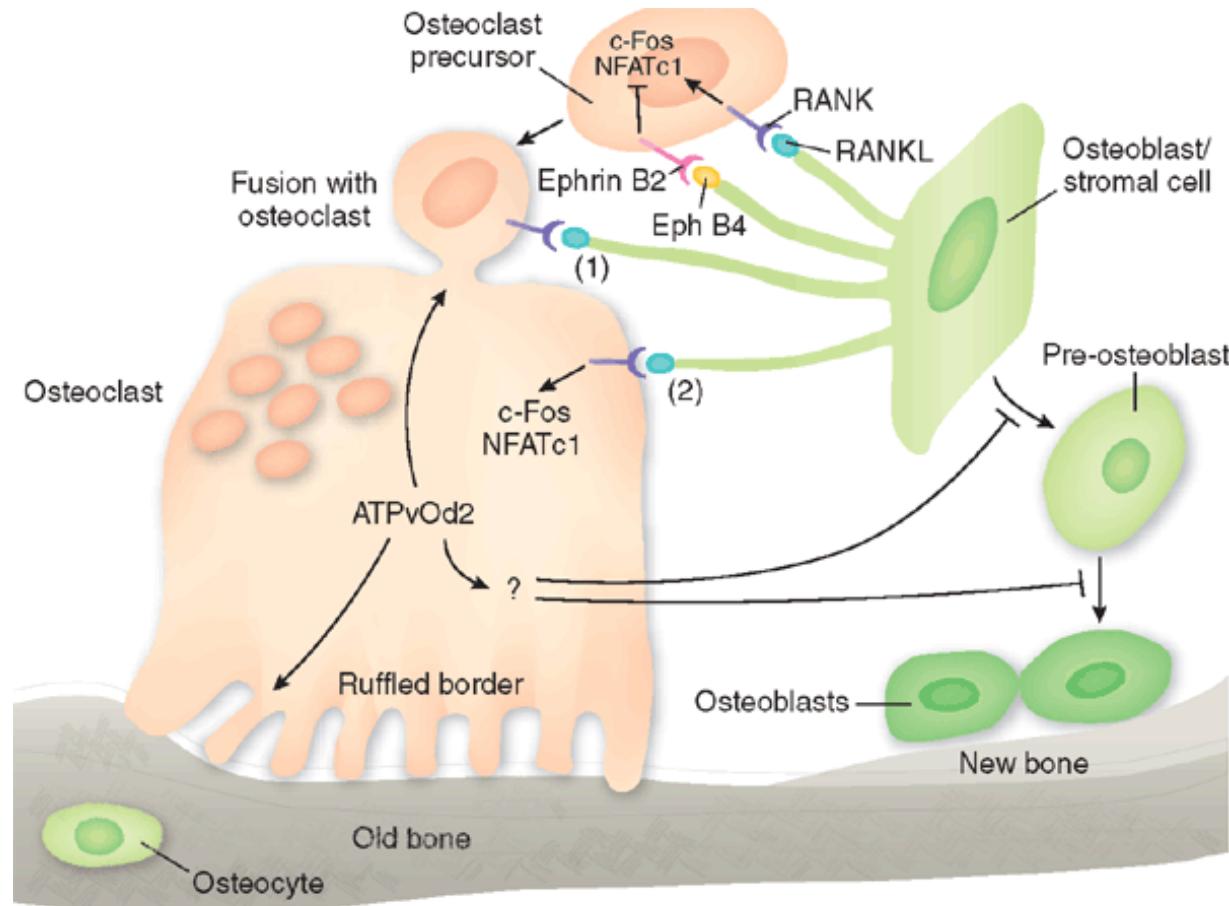
Identifying a disease target

1. Identify disease

- Osteoporosis (for example)

2. Identify a disease target for a drug

- What is mechanism of osteoporosis?
 - Osteoblasts are bone-forming cells
 - Osteoclasts are bone-degrading cells
 - Via bone mineral resorption
- In mid-to-late adulthood, balance in bone function shifts towards osteoclasts
 - Results in net loss of bone mineral density
 - More porous bones --> greater risk of fractures and breaks

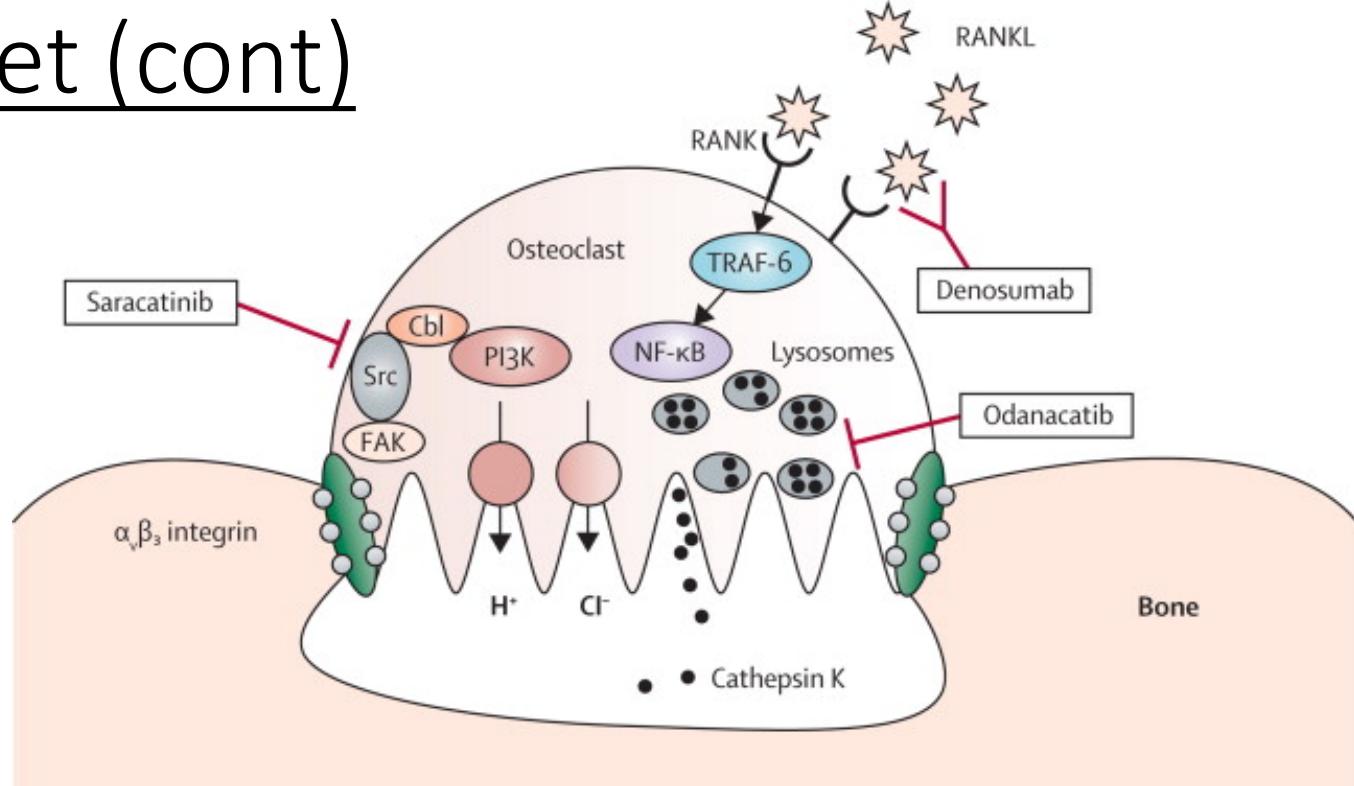


<http://www.nature.com/nm/journal/v12/n12/images/nm1206-1356-F1.gif>

Identifying a disease target (cont)

- Osteoclast function

- Cleavage of a peptide crosslink to type I collagen in bone during resorption process
 - Results in elevated carboxy-terminal portion of that peptide crosslink (CTX) in urine and serum
 - Urine/serum levels of CTX provide diagnostic measure of osteoclast activity
 - Urinary CTX (uCTX) is therefore an indirect pharmacodynamic (PD) biomarker
 - Would change despite which mechanism of osteoclast function was altered



<http://www.thelancet.com/cms/attachment/2001011910/2003800902/gr2.jpg>

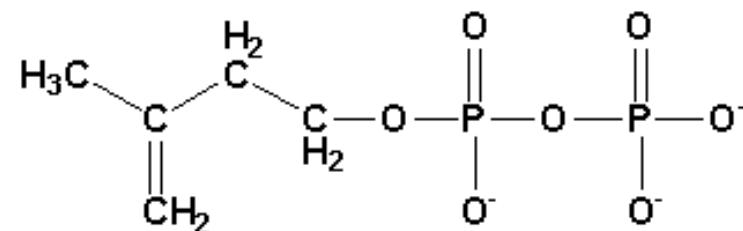
3. Lead Compound Development

- Once know target, need (ideally) a 3-D crystal structure of target protein
 - Can map the binding site (if target a protein receptor)
 - If target is ligand activated (or inactivated), compound development around ligand molecular structure
 - In our example, need an inhibitor of some mechanism of osteoclasts
 - Farnesyl diphosphate synthase (FPPS)
 - Inhibition of this enzyme disrupts the HMG-CoA reductase pathway, leading to apoptosis
 - Other compounds (e.g. statins) can inhibit HMG-CoA pathway, but statins do not adsorb to bone
 - Bisphosphonates selectivity to bone makes them ideal for osteoclasts/osteoporosis

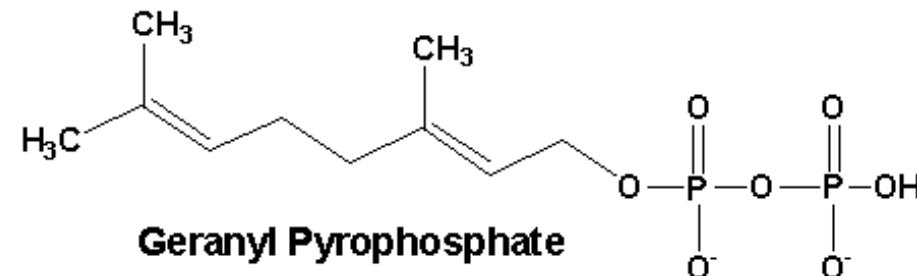
Lead Compound Development (cont)

- In this example, target is farnesyl diphosphate synthase (FPPS)

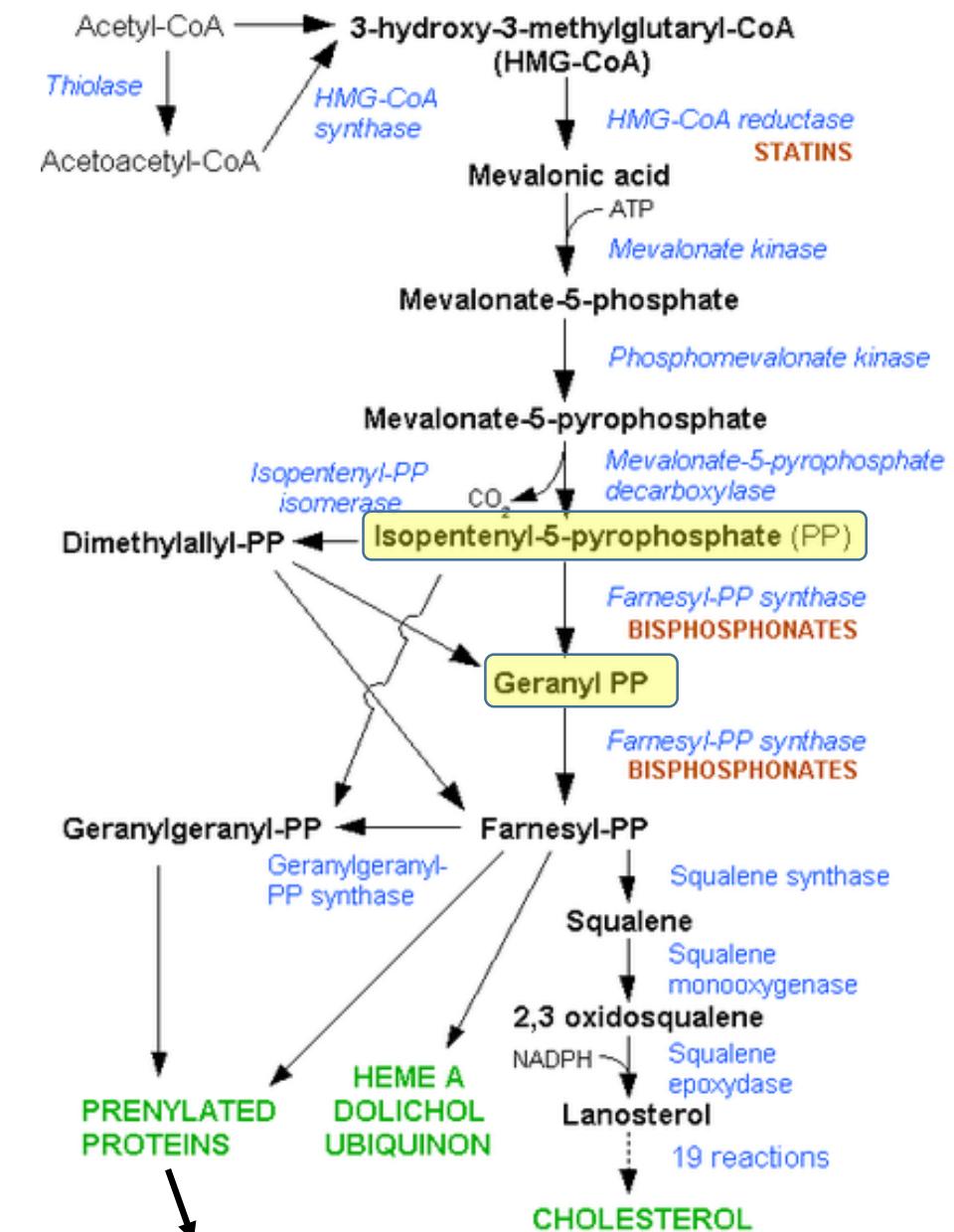
- Endogenous ligands are:



Isopentenyl Pyrophosphate



Geranyl Pyrophosphate



*Necessary for cell membrane integrity

Lead Compound Development (cont)

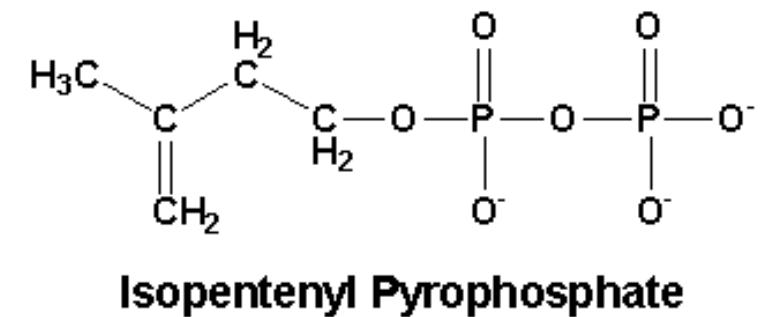
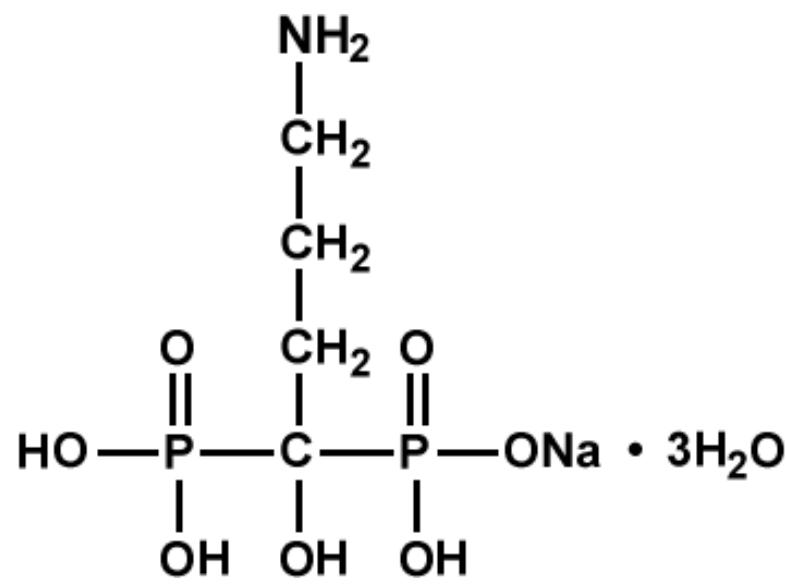
- Lead compound (for targeted therapy) must resemble endogenous ligand(s) in order to have similar affinity to agonize or antagonize target (FPPS)
- Screen 10,000s of molecularly-modeled compounds
- Obtain a list of leads (<100 cmpds)
- Test leads *in vitro* to ensure will hit target (e.g. FPPS) and induce theorized effect (e.g. induce osteoclast apoptosis via FPPS inhibition)
- Take leads (<5 cmpds) into **animals** for efficacy, safety, and other pharmacology experiments

Lead Compound Development (cont)

- Lipinski's Rule of Five
 - Compounds with two or more of the following are likely to demonstrate poor bioavailability (oral absorption, cell permeation):
 1. More than **five** hydrogen bond donors
 2. Molecular weight > 500 daltons
 3. Log P (permeability) > 5
 4. Sum of N's and O's (rough measure of H-bond acceptors) > 10
 - this is a general rule, not gospel

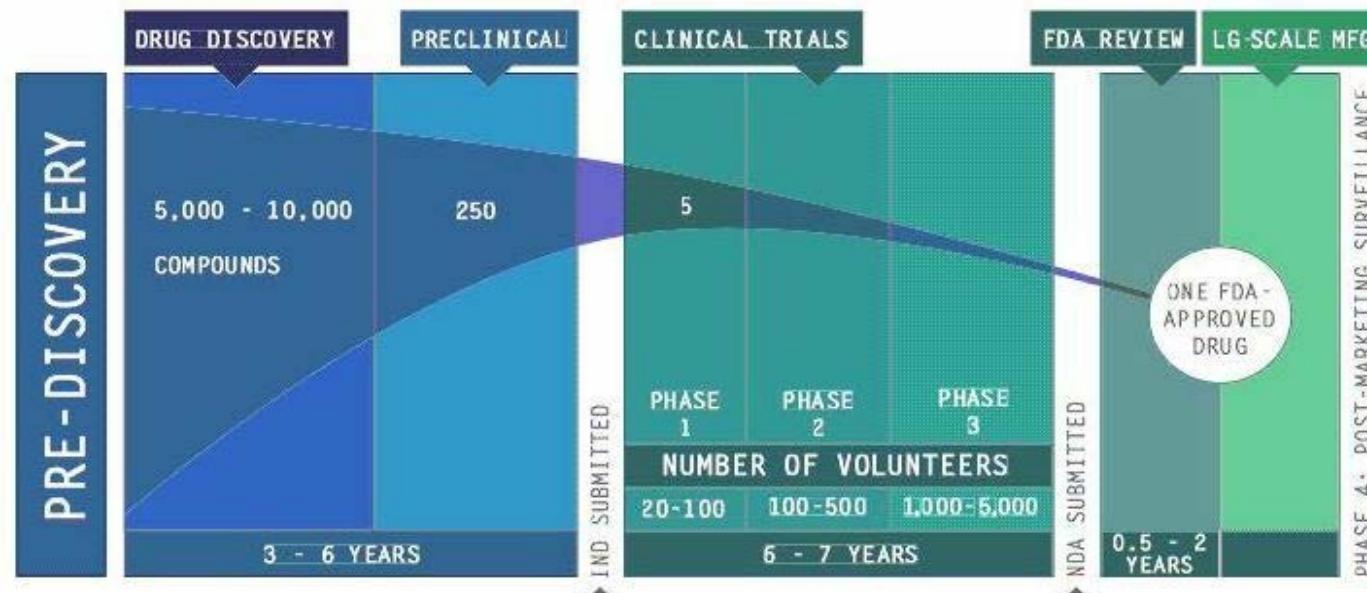
Lead Compound Development (cont)

- **Alendronate**, for example, emerged initial screening as *the* lead cmpd
 - Advances to clinical studies



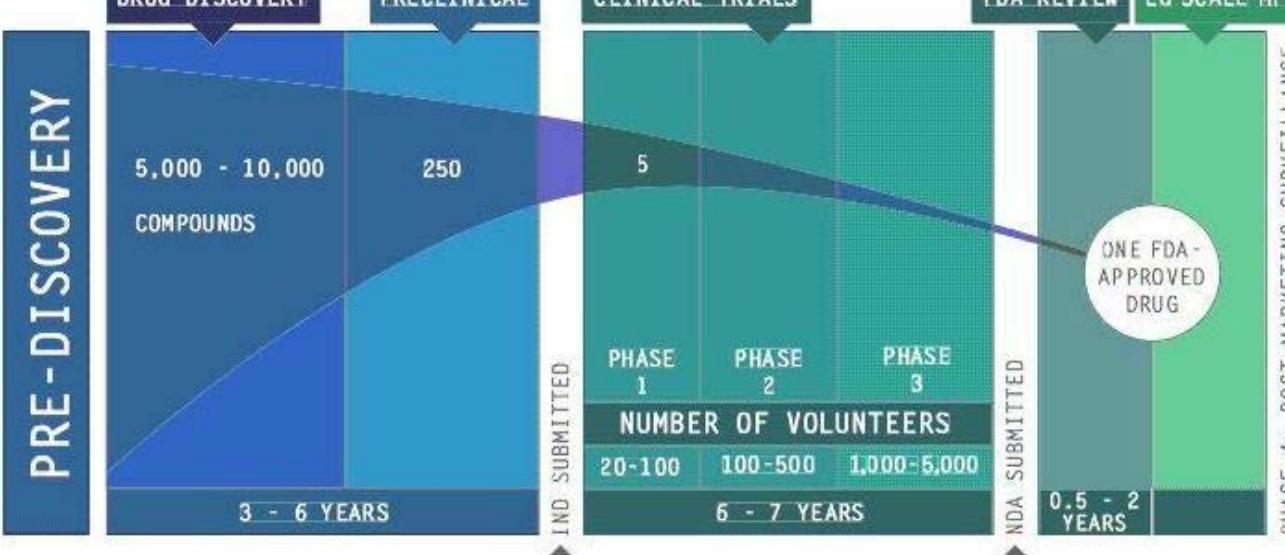
Clinical Trials

- Purpose of clinical trials is to demonstrate a drug is effective and safe
 - Early phases to optimize dose, frequency, and route (via pharmacokinetics)
 - Latter phases to ensure safe and effective therapy and to justify optimized dose
- FDA reviews all clinical data to verify exposure-response relationships (pharmacology)
 - A given dose of drug will expose subject to a given concentration of drug for a given amount of time
 - That measured drug concentration in plasma will lead a quantifiable pharmacological response, measured by some biomarker (e.g. uCTX)



Reasons for Early-Stage Slowdowns

- Synthetic complexity
- Lack of *in vitro* efficacy
- Stability issues



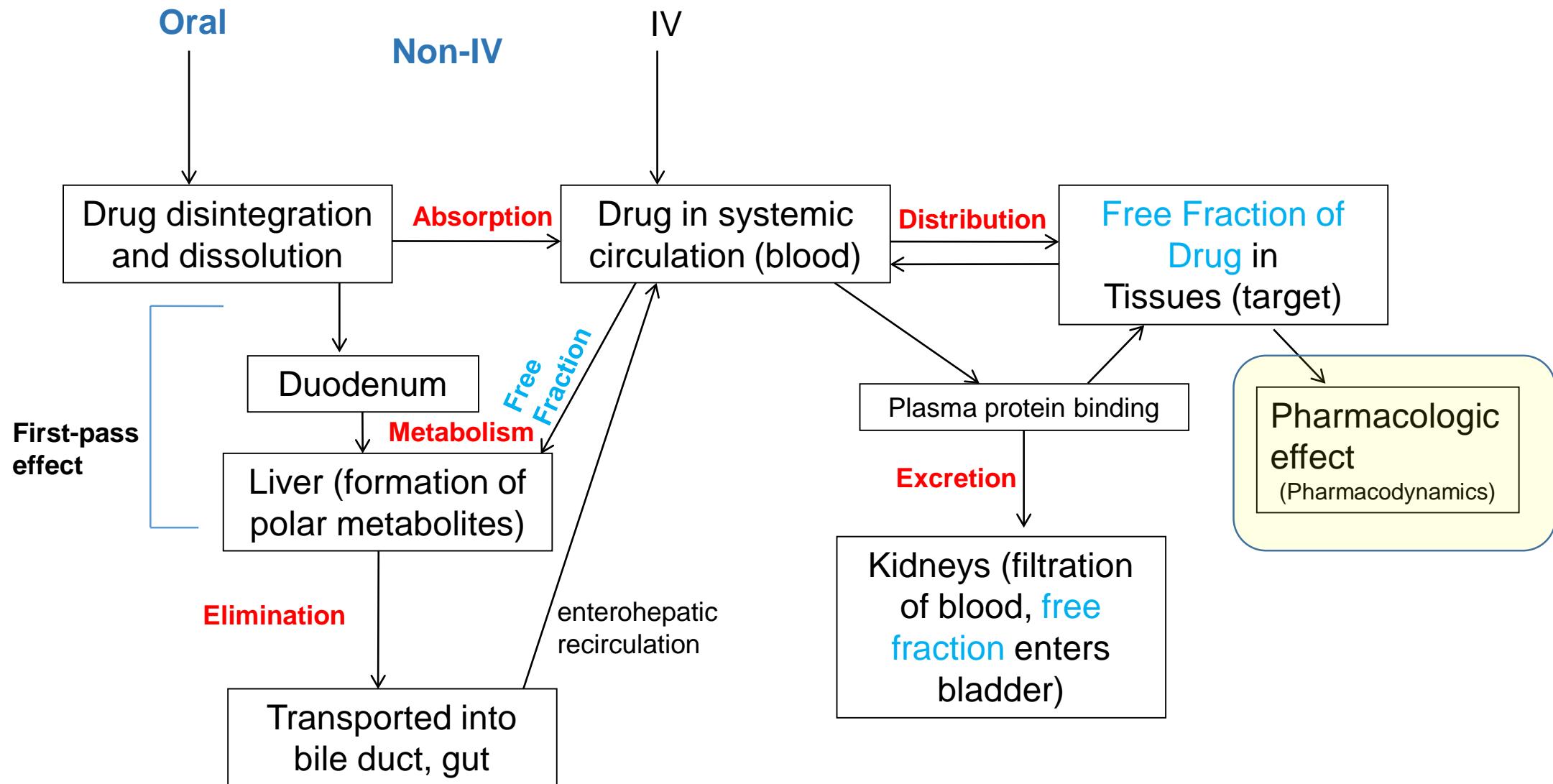
Reasons for Late-Stage Failures

- Toxicity
- Lack of clinical efficacy
- Market reasons
- Poor biopharmaceutical properties
 - E.g. poor bioavailability, instability, short half-life

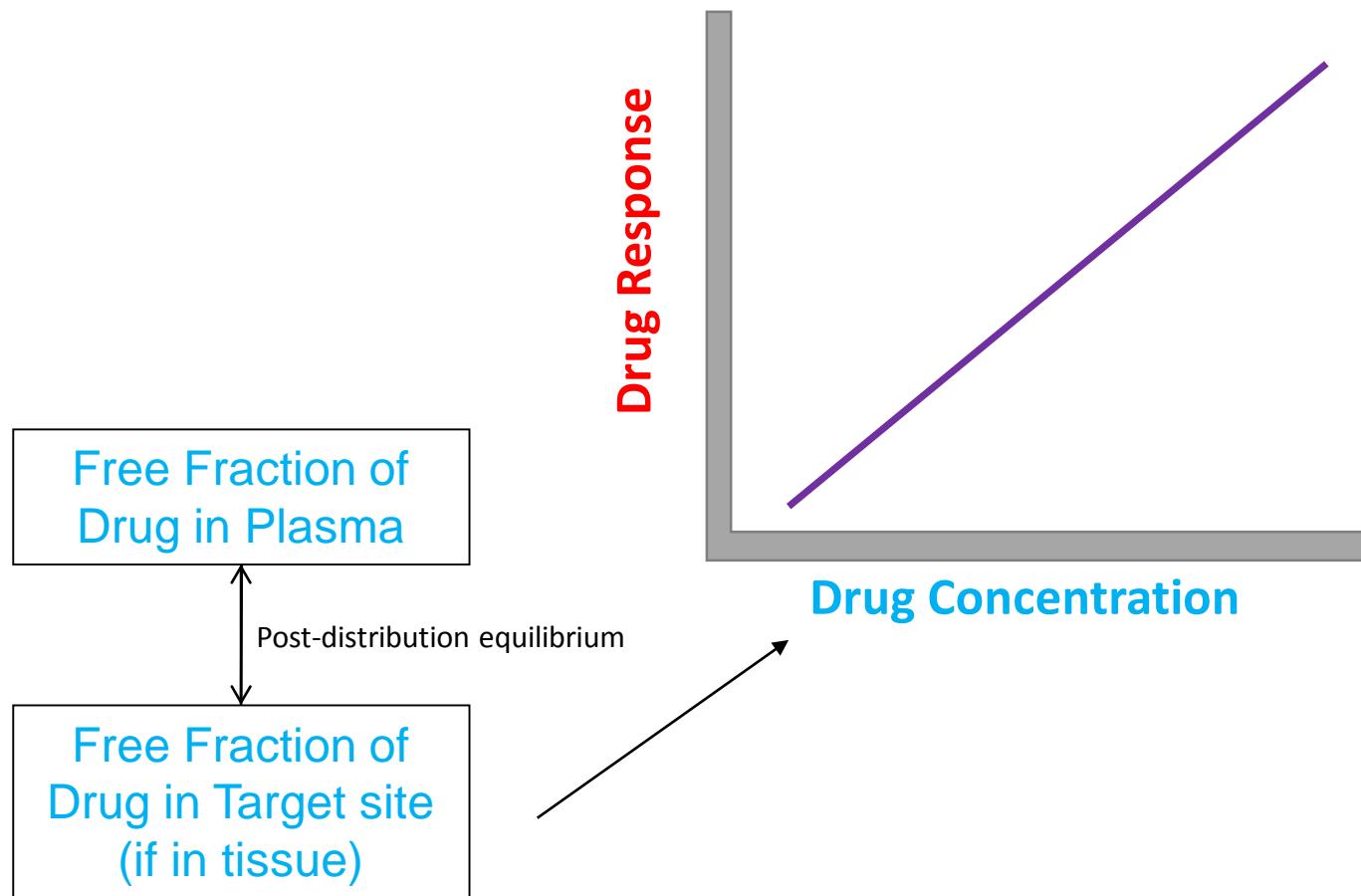
Pharmacology is Ultimately What's Reviewed

- **Pharmacokinetics** (what the body does to the drug)
 - Physicochemical properties of the drug molecule (i.e. solubility, lipophilicity, stability)
 - Absorption, Distribution, Metabolism, Elimination (ADME)
 - Understand kinetic rates to optimize dose amount, frequency, and adjustments needed in subpopulations (age, race, body size, genotype, etc)
 - Can develop mathematical models around observed data
 - Used to simulate altered dose/regimen to predict outcome of trials before run
 - Particularly useful in neonates, pediatrics, and geriatrics, where clinical trials difficult
- **Pharmacodynamics** (what the drug does to the body)
 - Understanding biological target and any off-target effects (aka side effects)
 - Location, structure, expression, associated entities, etc
 - Mechanism of action in normal and disease states

Pharmacokinetics (ADME) flow chart



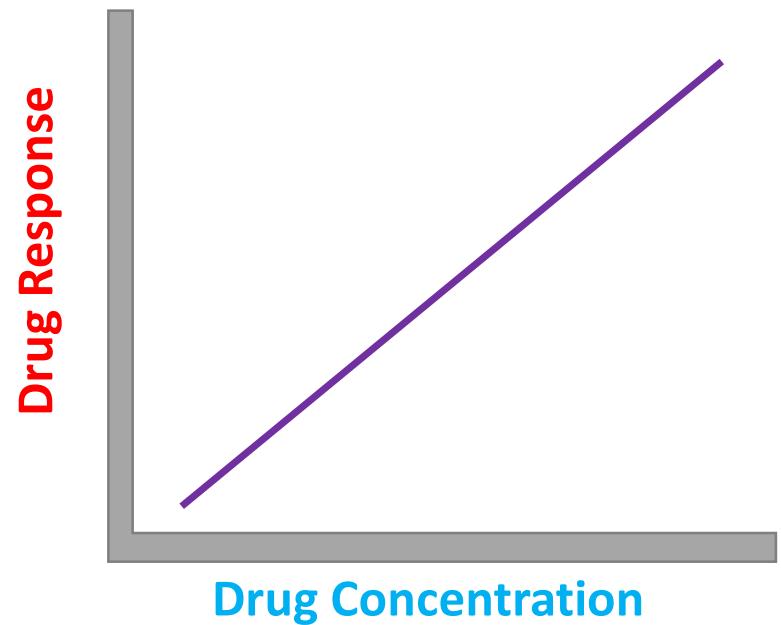
Pharmacodynamics flow chart



- Biomarker relevant to disease
- **Response based on mechanism of drug effect**
- Drug effect could be:
 - Direct stimulatory (e.g. albuterol)
 - Direct inhibitory (e.g. Plavix®)
 - Indirect stimulatory (e.g. SSRIs)
 - Indirect inhibitory (e.g. bisphosphonates)

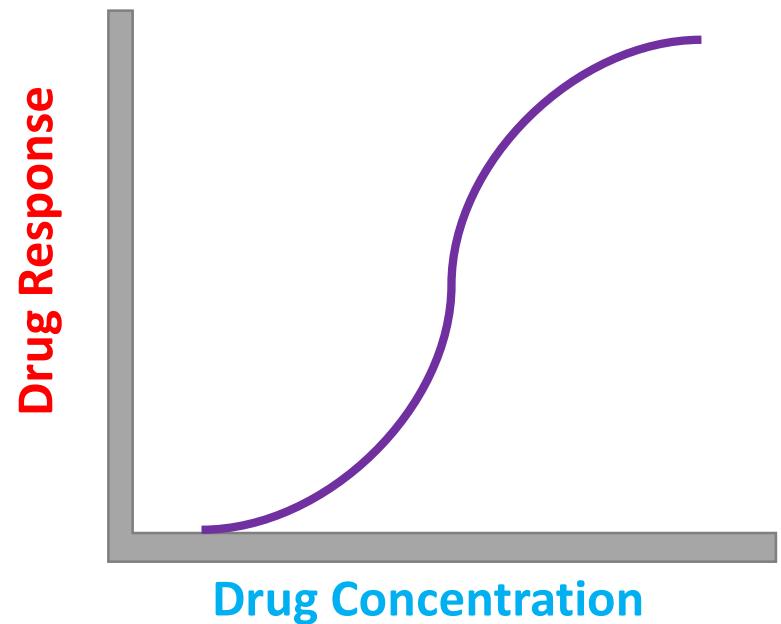
Exposure/Response Analyses

- The shape of the exposure/response curve is dependent on:
 1. Disease target
 2. Drug pharmacology
 3. Type of Response being monitored
 - Pharmacological endpoints
 - Toxicological endpoints
- Not always linear...
 - Could be plateau, semi-parabolic, semi-hyperbolic



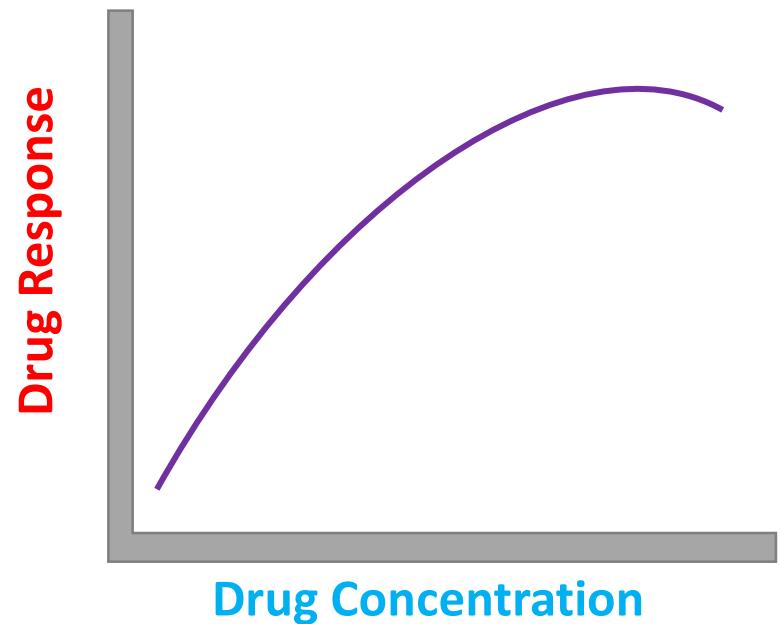
Exposure/Response Analyses

- The shape of the exposure/response curve is dependent on:
 1. Disease target
 2. Drug pharmacology
 3. Type of Response being monitored
 - Pharmacological endpoints
 - Toxicological endpoints
- Not always linear...
 - Could be plateau, semi-parabolic, semi-hyperbolic



Exposure/Response Analyses

- The shape of the exposure/response curve is dependent on:
 1. Disease target
 2. Drug pharmacology
 3. Type of Response being monitored
 - Pharmacological endpoints
 - Toxicological endpoints
- Not always linear...
 - Could be plateau, semi-parabolic, semi-hyperbolic



Exposure/Response Analyses

- Requires fundamental understanding of:

Pharmacokinetics (exposure)

- Chemistry (organic, inorganic, biochem, analytical, physical)
- Biology
- Mathematics

Pharmacodynamics (response)

- Physiology
- Anatomy
- Molecular biology
- Genetics

Exposure/Response Analyses (cont)

- Requires extensive education (usually MS or PhD)
- Usually requires expensive proprietary software
 - Training on the software usually obtained via company-guided workshops for an additional fee
- R is an open-source, free software program
 - Pros
 - Performs basic and complex mathematical operations
 - Input outside datasets
 - Develop models, plots, etc
 - Can simulate data based on developed models
 - Cons
 - Requires computer programming skills, training
 - Requires intermediate understanding of statistics (otherwise, functions won't run)
 - There's no buttons to click to perform stats, everything entered manually

Applications for Exposure/Response Modeling

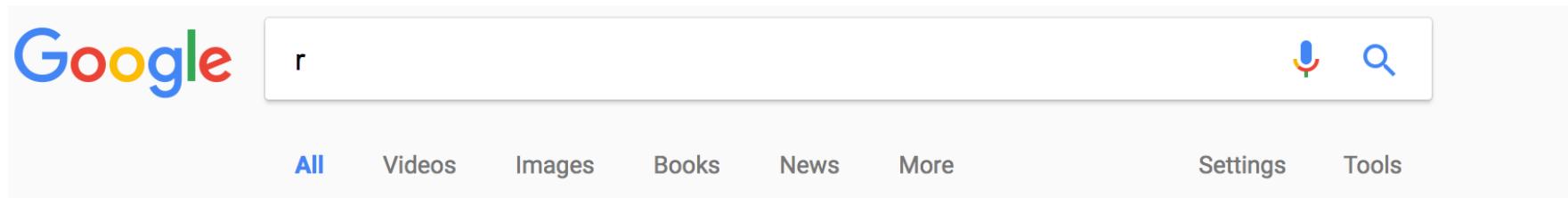
- If can predict what dose causes what exposure, and what exposure causes what response...
 - ... then can give a dose and predict responses
 - Both drug's active pharmacological effect and side effects
- Can apply more personalized doses to patients
 - If genetic polymorphisms in drug metabolizing enzymes and transporters are affecting drug clearance, hence drug exposure, then altering the level of response those patients have to the drug
 - Can alter dose for those patients so they receive a “normal” amount of exposure to induce a normal amount of response.

Course Objectives

- To teach non-pharmacometrists the basic theory of exposure/response analyses
- To introduce R and its ability to conduct exploratory exposure/response analyses
- To instill an appreciation of the amount of data, time, and energy needed to get a drug approved

Installing R

- Go to <http://cran.r-project.org/>
 - Available for Windows 7, 8, 10 or OS X
 - When downloading for the first time, install “base”
 - As of April 21, 2017, version 3.3.3 is newest



The Comprehensive R Archive Network
<https://cran.r-project.org/> ▾
Download and Install R. Precompiled binary distributions of the base system and contributed packages, Windows and Mac users most likely want one of these ...
Packages · CRAN - Mirrors · Manuals · Task Views

R: The R Project for Statistical Computing
<https://www.r-project.org/> ▾
R, also called GNU S, is a strongly functional language and environment to statistically explore data sets, make many graphical displays of data from custom ...
CRAN - Mirrors · Of /src/base/R-3 · R: What is R? · Manuals

• Install based on your operating system



[CRAN
Mirrors](#)
[What's new?](#)
[Task Views](#)
[Search](#)

[About R](#)
[R Homepage](#)
[The R Journal](#)

[Software](#)
[R Sources](#)
[R Binaries](#)
[Packages](#)
[Other](#)

[Documentation](#)
[Manuals](#)
[FAQs](#)
[Contributed](#)

The Comprehensive R Archive Network

Download and Install R

Precompiled binary distributions of the base system and contributed packages. **Windows** and **Mac** users most likely want one of these versions of R:

- [Download R for Linux](#)
- [Download R for \(Mac\) OS X](#)
- [Download R for Windows](#)

R is part of many Linux distributions, you should check with your Linux package management system in addition to the link above.

Source Code for all Platforms

Windows and Mac users most likely want to download the precompiled binaries listed in the upper box, not the source code. The sources have to be compiled before you can use them. If you do not know what this means, you probably do not want to do it!

- The latest release (Monday 2017-03-06, Another Canoe) [R-3.3.3.tar.gz](#), read [what's new](#) in the latest version.
- Sources of [R alpha and beta releases](#) (daily snapshots, created only in time periods before a planned release).
- Daily snapshots of current patched and development versions are [available here](#). Please read about [new features and bug fixes](#) before filing corresponding feature requests or bug reports.
- Source code of older versions of R is [available here](#).
- Contributed extension [packages](#)

Questions About R

- If you have questions about R like how to download and install the software, or what the license terms are, please read our [answers to frequently asked questions](#) before you send an email.

What are R and CRAN?

R is ‘GNU S’, a freely available language and environment for statistical computing and graphics which provides a wide variety of statistical and graphical techniques: linear and nonlinear modelling, statistical tests, time series analysis, classification, clustering, etc. Please consult the [R project homepage](#) for further information.

CRAN is a network of ftp and web servers around the world that store identical, up-to-date, versions of code and documentation for R. Please use the CRAN [mirror](#) nearest to you to minimize network traffic.

Submitting to CRAN

To “submit” a package to CRAN, check that your submission meets the [CRAN Repository Policy](#) and then use the [web form](#).

If this fails, upload to <ftp://CRAN.R-project.org/incoming/pretest> and send an email to CRAN@R-project.org following the policy. Please do not attach submissions to emails, because this will clutter mailboxes of half a dozen people.

Note that we generally do not accept submissions of precompiled binaries due to security reasons. All binary distribution listed above are compiled by selected maintainers, who are in charge for all of their platform, respectively.



[CRAN](#)
[Mirrors](#)
[What's new?](#)
[Task Views](#)
[Search](#)

[About R](#)
[R Homepage](#)
[The R Journal](#)

[Software](#)
[R Sources](#)
[R Binaries](#)
[Packages](#)
[Other](#)

[Documentation](#)
[Manuals](#)
[FAQs](#)
[Contributed](#)

R for Windows

Subdirectories:

[base](#)

Binaries for base distribution (managed by Duncan Murdoch). This is what you want to [install R for the first time](#).

[contrib](#)

Binaries of contributed CRAN packages (for R \geq 2.11.x; managed by Uwe Ligges). There is also information on [third party software](#) available for CRAN Windows services and corresponding environment and make variables.

[old contrib](#)

Binaries of contributed CRAN packages for outdated versions of R (for R $<$ 2.11.x; managed by Uwe Ligges).

[Rtools](#)

Tools to build R and R packages (managed by Duncan Murdoch). This is what you want to build your own packages on Windows, or to build R itself.

Please do not submit binaries to CRAN. Package developers might want to contact Duncan Murdoch or Uwe Ligges directly in case of questions / suggestions related to Windows binaries.

You may also want to read the [R FAQ](#) and [R for Windows FAQ](#).

Note: CRAN does some checks on these binaries for viruses, but cannot give guarantees. Use the normal precautions with downloaded executables.

- Click link for R-3.3.3
- Save to desktop
- Open and download (make sure you have Admin rights to your computer)



[CRAN](#)
[Mirrors](#)
[What's new?](#)
[Task Views](#)
[Search](#)

[About R](#)
[R Homepage](#)
[The R Journal](#)

[Software](#)
[R Sources](#)
[R Binaries](#)
[Packages](#)
[Other](#)

[Documentation](#)
[Manuals](#)
[FAQs](#)
[Contributed](#)

R-3.3.3 for Windows (32/64 bit)

[Download R 3.3.3 for Windows](#) (71 megabytes, 32/64 bit)
[Installation and other instructions](#)
[New features in this version](#)

If you want to double-check that the package you have downloaded matches the package distributed by CRAN, you can compare the [md5sum](#) of the .exe to the [fingerprint](#) on the master server. You will need a version of md5sum for windows: both [graphical](#) and [command line versions](#) are available.

Frequently asked questions

- [Does R run under my version of Windows?](#)
- [How do I update packages in my previous version of R?](#)
- [Should I run 32-bit or 64-bit R?](#)

Please see the [R FAQ](#) for general information about R and the [R Windows FAQ](#) for Windows-specific information.

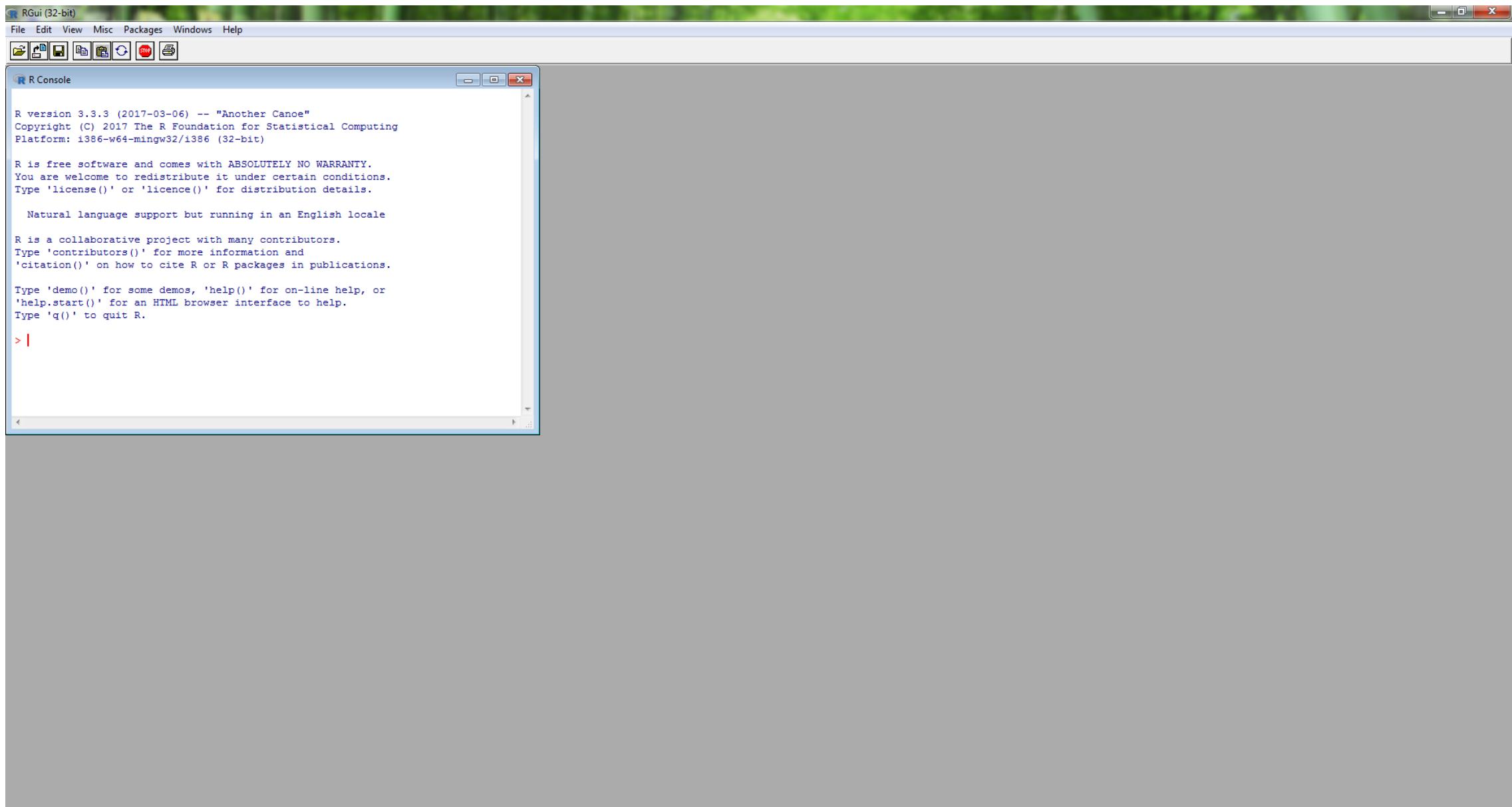
Other builds

- Daily alpha/beta/rc [builds of the upcoming R 3.4.0](#).
- Patches to this release are incorporated in the [r-patched snapshot build](#).
- A build of the development version (which will eventually become the next major release of R) is available in the [r-devel snapshot build](#).
- [Previous releases](#)

Note to webmasters: A stable link which will redirect to the current Windows binary release is
<CRAN MIRROR>/bin/windows/base/release.htm

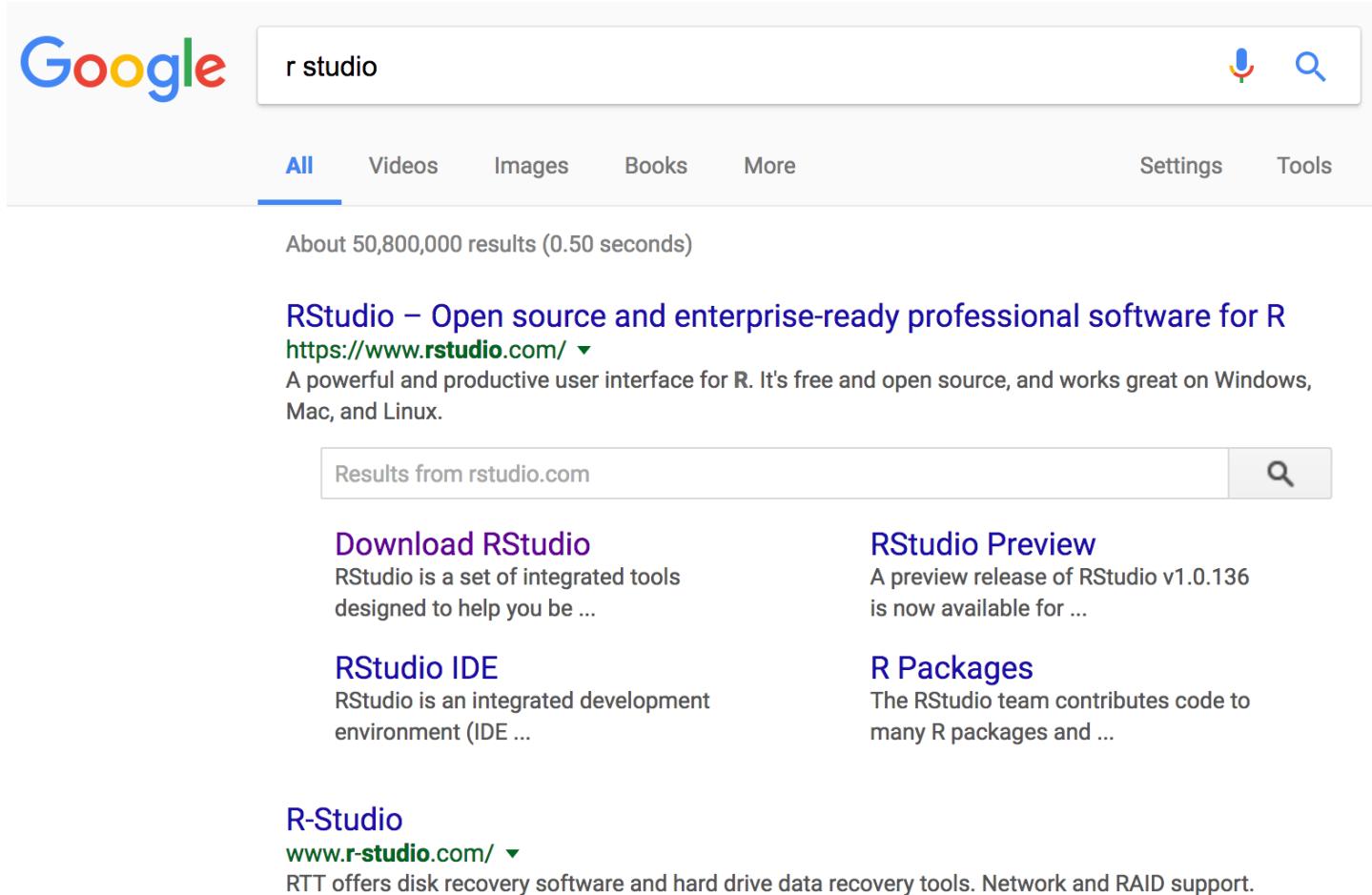
Last change: 2017-03-06, by Duncan Murdoch

- Open R to make sure it's working



Installing R Studio

- Once R installed on Desktop, go to www.rstudio.com



A screenshot of a Google search results page for the query "r studio". The search bar shows the query. Below it, the "All" tab is selected, along with other options like Videos, Images, Books, and More. The search results show approximately 50,800,000 results found in 0.50 seconds. The top result is a link to the official RStudio website, which is described as "Open source and enterprise-ready professional software for R". The description below the link states: "A powerful and productive user interface for R. It's free and open source, and works great on Windows, Mac, and Linux." Below the main search results, there are two sections: "Results from rstudio.com" and "Results from r-studio.com/". The first section contains links to "Download RStudio" (described as a set of integrated tools), "RStudio IDE" (described as an integrated development environment), and "R Packages" (described as the team contributing code to many R packages). The second section contains a link to "R-Studio" (described as offering disk recovery software and hard drive data recovery tools, with Network and RAID support).

Google r studio

All Videos Images Books More Settings Tools

About 50,800,000 results (0.50 seconds)

RStudio – Open source and enterprise-ready professional software for R
<https://www.rstudio.com/> ▾

A powerful and productive user interface for R. It's free and open source, and works great on Windows, Mac, and Linux.

Results from rstudio.com

Download RStudio
RStudio is a set of integrated tools designed to help you be ...

RStudio IDE
RStudio is an integrated development environment (IDE ...)

R Packages
The RStudio team contributes code to many R packages and ...

R-Studio
www.r-studio.com/ ▾

RTT offers disk recovery software and hard drive data recovery tools. Network and RAID support.

- Products → RStudio

The screenshot shows the RStudio website homepage. The header includes the RStudio logo, navigation links for 'rstudio::conf', 'Products', 'Resources', 'Pricing', 'About Us', 'Blogs', and a search icon. A prominent banner on the left features the text 'RStudio' in large white letters, 'Open source and enterprise-grade professional software for R', and a yellow exclamation mark icon with the text 'Announcing RStudio'. To the right of the banner, a sidebar menu is open under the 'Products' category, listing 'RStudio', 'Shiny', 'R Packages', 'RStudio Server Pro', 'RStudio Connect', 'Shiny Server Pro', and 'shinyapps.io'. The background of the page has a blue gradient with faint text: 'Download RStudio', 'Discover Shiny', 'shinyapps.io Login', and 'Discover Connect'.

- RStudio
- rstudio::conf
- Products
- Resources
- Pricing
- About Us
- Blogs
- Q

RStudio

Shiny

R Packages

RStudio Server Pro

RStudio Connect

Shiny Server Pro

shinyapps.io

! Announcing RStudio

Open source and enterprise-grade professional software for R

Download RStudio

Discover Shiny

shinyapps.io Login

Discover Connect

● ● ● ●

- Download Desktop version

 R Studio

rstudio::conf

Products Resources Pricing About Us Blogs 

RStudio is an integrated development environment (IDE) for R. It includes a console, syntax-highlighting editor that supports direct code execution, as well as tools for plotting, history, debugging and workspace management. [Click here to see more RStudio features.](#)

RStudio is available in open source and commercial editions and runs on the desktop (Windows, Mac, and Linux) or in a browser connected to RStudio Server or RStudio Server Pro (Debian/Ubuntu, RedHat/CentOS, and SUSE Linux).



Desktop

Run RStudio on your desktop

[RStudio > Desktop](#)



Server

Centralize access and computation

[RStudio Server >](#)



01:31

HD



[CLICK HERE TO SEE ADDITIONAL FEATURES](#)

- Follow links to download RStudio

The screenshot shows the RStudio website's Overview page. At the top, there's a navigation bar with links for Products, Resources, Pricing, About Us, Blogs, and a search icon. Below the navigation, there's a section titled "What is RStudio?" with a bulleted list of features. Further down, there are sections for "Support" (Community forums only) and "License" (AGPL v3). At the bottom, there are two prominent blue buttons: "DOWNLOAD RSTUDIO DESKTOP" and "BUY NOW".

R Studio

Products Resources Pricing About Us Blogs

What is RStudio?

- Syntax highlighting, code completion, and smart indentation
- Execute R code directly from the source editor
- Quickly jump to function definitions
- Easily manage multiple working directories using projects
- Integrated R help and documentation
- Interactive debugger to diagnose and fix errors quickly
- Extensive package development tools

Overview

Support Community forums only

- Priority Email Support
- 8 hour response during business hours (ET)

License AGPL v3

[RStudio License Agreement](#)

Pricing Free

\$995/year

[DOWNLOAD RSTUDIO DESKTOP](#)

[BUY NOW](#)

- Download based on your operating system. Save to Desktop.



rstudio::conf

Products

Resources

Pricing

About Us

Blogs



RStudio is a set of integrated tools designed to help you be more productive with R. It includes a console, syntax-highlighting editor that supports direct code execution, as well as tools for plotting, history, debugging and workspace management.

If you run R on a Linux server and want to enable users to remotely access RStudio using a web browser [please download RStudio Server](#).

Do you need support or a commercial license? [Check out our commercial offerings](#)

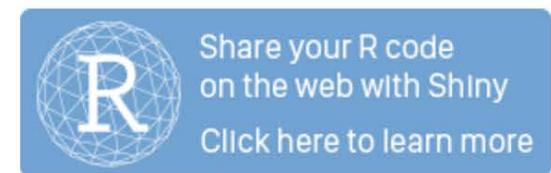
RStudio Desktop 1.0.44 — Release Notes

RStudio requires R 2.11.1 (or higher). If you don't already have R, you can download it [here](#).

Installers for Supported Platforms

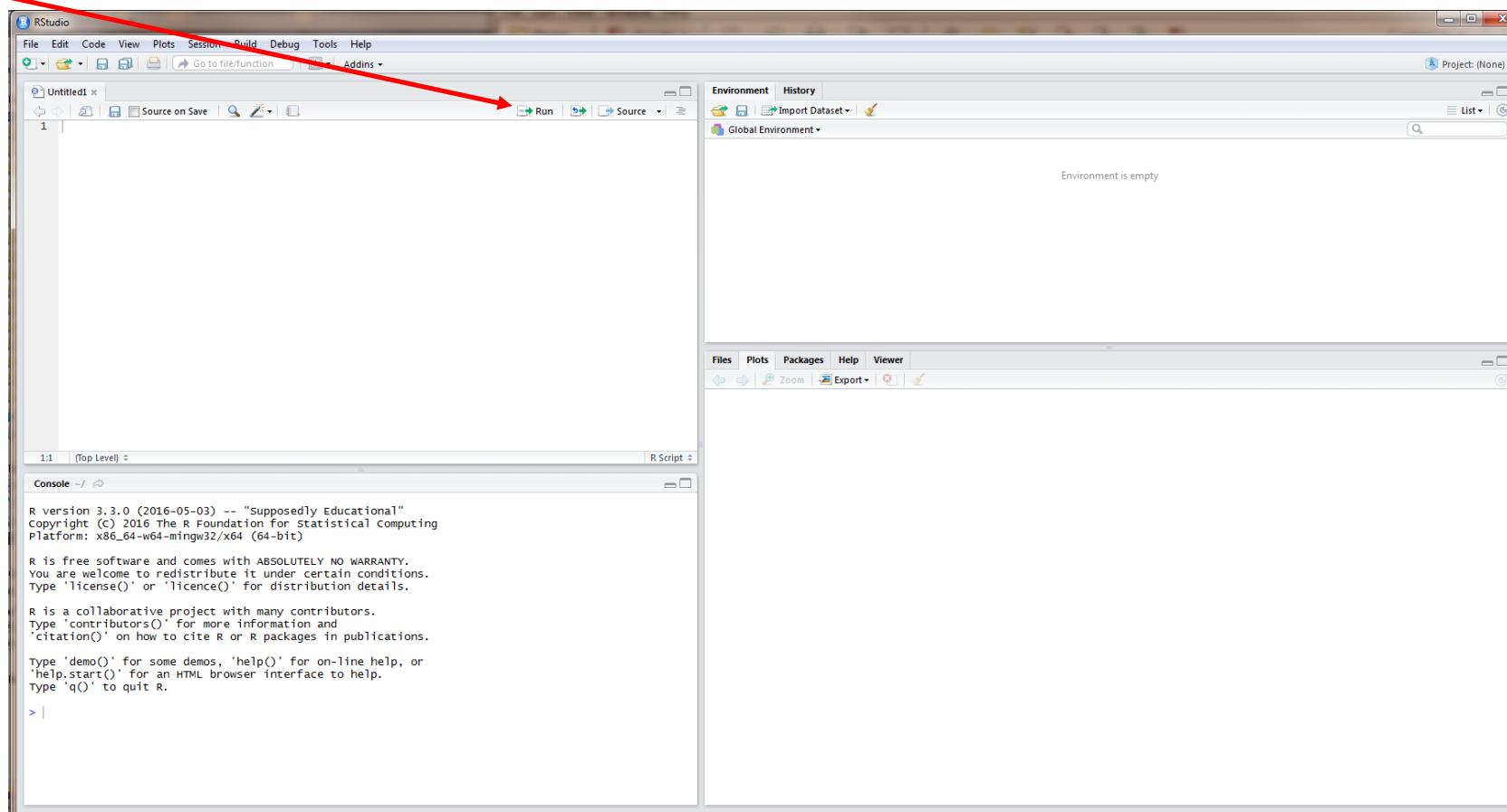
Installers

RStudio 1.0.44 - Windows Vista/7/8/10	81.9 MB	2016-11-01	7ccedc36c1f0a0861393763cfbe1c61d
RStudio 1.0.44 - Mac OS X 10.6+ (64-bit)	71.1 MB	2016-11-01	32256c7ac6d6597192a1bafa56a2747f
RStudio 1.0.44 - Ubuntu 12.04+/Debian 8+ (32-bit)	85.4 MB	2016-11-01	5f7fb95ee727606e9779af7bfe6fc6a8
RStudio 1.0.44 - Ubuntu 12.04+/Debian 8+ (64-bit)	92 MB	2016-11-01	074b7d3336ad07e32d10553f9669194a
RStudio 1.0.44 - Fedora 19+/RedHat 7+/openSUSE 13.1+ (32-bit)	84.6 MB	2016-11-01	a5b203d482c6ab9ab77c5daf3fad5b8a
RStudio 1.0.44 - Fedora 19+/RedHat 7+/openSUSE 13.1+ (64-bit)	85.6 MB	2016-11-01	bdc2cf31061d393a5d6626329f19bd6f



Workings of R Studio

- Use “#” to write a comment without R thinking it is code or a command
- After coding, execute using “Ctl+R” (on PC), “Cmd+Enter” (on Mac), or clicking “Run”



Interacting with RStudio

The screenshot illustrates the RStudio interface with four main panes:

- Pane 1 (Source):** Displays the `labike` dataset as a table with 38 observations and 6 variables: name, longitude, latitude, type, bike_count_pm, and ped_count. The variable `name` lists locations like "1st & Alameda", "4th & Wilton", etc.
- Pane 2 (Environment etc.):** Shows the `labike` dataset in the Workspace pane, described as "38 obs. of 6 variables".
- Pane 3 (Files etc.):** Lists files in the current directory, including `.Rprofile`, `bus_stops_df.rda`, `captions.txt`, `CATwitter.robject`, `cdc.rda`, `labike.csv`, `NJTwitter.robject`, a folder named `R`, `smallcaptions.robject`, `survey.rda`, `twitterwithdate.csv`, and `weather.robject`.
- Pane 4 (Console):** Displays the R startup message, the `labike` dataset loading command, and the `View(labike)` command being entered.

R: Basic concepts

- Objects in R
 - 1-dimensional
 - 2-dimensional
 - Atomic
 - Recursive
 - Subsetting
 - []
 - \$
 - Names
 - Adding/Removing Rows/Columns
- Loading Datasets
 - Setting the working directory
 - `read.csv()`
- Install packages
 - `install.packages()`
 - Packages tab

Objects in R

- Let's create our first object in R by typing the following into the console
“val <- 3”
- When you hit enter you will see that val appears in the upper right pane.
- We have created a **value**, an object with a single data element, but objects can contain many data elements.

One-dimensional Objects

- **Vectors** are **atomic** objects i.e. they contain data elements that are all of the same class
- Make a vector called vec:

```
vec <- c(1, 2, 3, 4, 5, 6)
```

- **Lists** are **recursive** objects i.e. they can contain many different classes of objects
- Make a list called ls:

```
ls <- list(1, 2, 3, "a", "b", "c")
```

Two-dimensional Objects

- **Matrices** are **atomic** objects i.e. they contain data elements that are all of the same class
- Make a matrix called m:

```
m<- matrix(c(1,2,3,4,5,6), nrow =2)
```

- **Data frames** are **recursive** objects i.e. they can contain many different classes of objects.
- Make a data frame called df:

```
df <- data.frame(x = 1:3, y = c("a", "b", "c"))
```

Subsetting Objects

There are three main ways to subset objects:

1. `df[2]` or `df["y"]`
2. `df[[2]]` or `df[["y"]]`
3. `df$y`

- The `[]` method can return multiple objects with names included
- `[[]]` and `$` both return single objects without names
 - `$` only works on recursive objects like lists and data frames
 - `$` will not work on atomic objects like vectors and matrices

Names

You can subset variables by row or column names.

Only the df object we created has names.

You can add names:

```
names(ls)<-c("A","B","C","D","E","F")
```

```
names(ls)<-LETTERS[1:6]
```

```
colnames(m)<- LETTERS[1:2]
```

```
rownames(m)<- LETTERS[24:26]
```

You can rename columns and rows

```
colnames(df)<- c("A","B")
```

```
colnames(df)<- LETTERS[1:2]
```

Adding columns

Add an unnamed third column to df and m

```
df[,3]<-1:3
```

```
m[,3]<-1:3
```

Add df to m as additional columns:

```
dfmc<-cbind(df,m)
```

cbind will only work if **row** names match!

Add a column named D to df

```
df$D<-1:4
```

This will not work for m!

Adding rows

Add an unnamed 4th row to df and m:

```
df[4,<-1:4
```

```
m[4,<-1:4
```

Add m to df as additional rows:

```
dfmr<-rbind(df,m)
```

rbind will only work if **column** names match!

Removing columns

- To delete the first column of m:

```
m <- m[, -1]
```

- To delete the B column of m:

```
m <- m[-B]
```

- To delete the first row of m:

```
m <- m[-1, ]
```

- To remove everything except column B in df:

```
df<-df$B
```

Introduction to Base Plotting in R

- Histogram
- Scatterplot
- Boxplot
- Parameters
- Additional functions
- Annotation
- Regression lines
- Multipanel plots

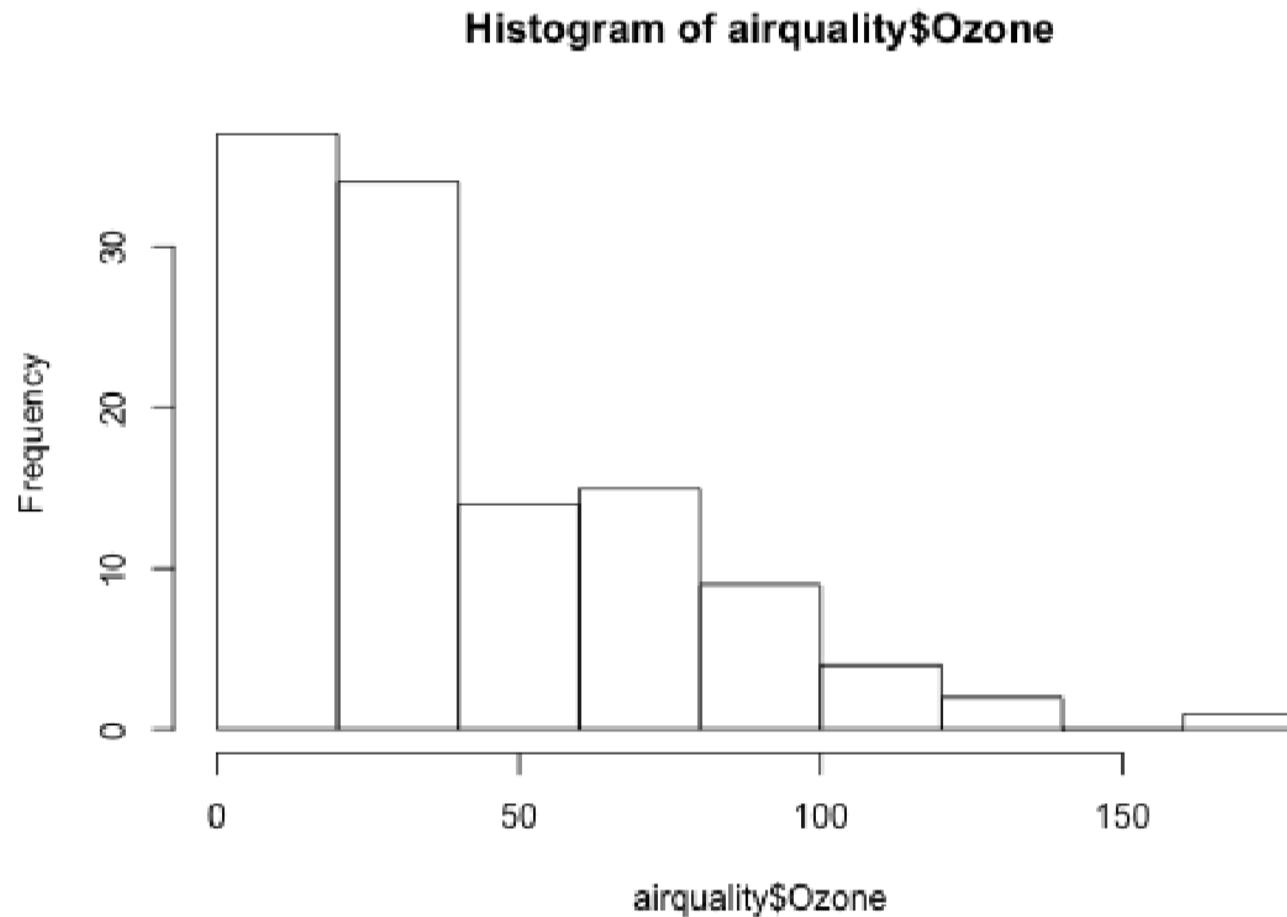
Base Graphics

Base graphics are used most commonly and are a very powerful system for creating 2-D graphics.

- There are two *phases* to creating a base plot
 - Initializing a new plot
 - Annotating (adding to) an existing plot
- Calling `plot(x, y)` or `hist(x)` will launch a graphics device (if one is not already open) and draw a new plot on the device
- If the arguments to `plot` are not of some special class, then the *default* method for `plot` is called; this function has *many* arguments, letting you set the title, x axis label, y axis label, etc.
- The base graphics system has *many* parameters that can be set and tweaked; these parameters are documented in `?par`; it wouldn't hurt to try to memorize this help page!

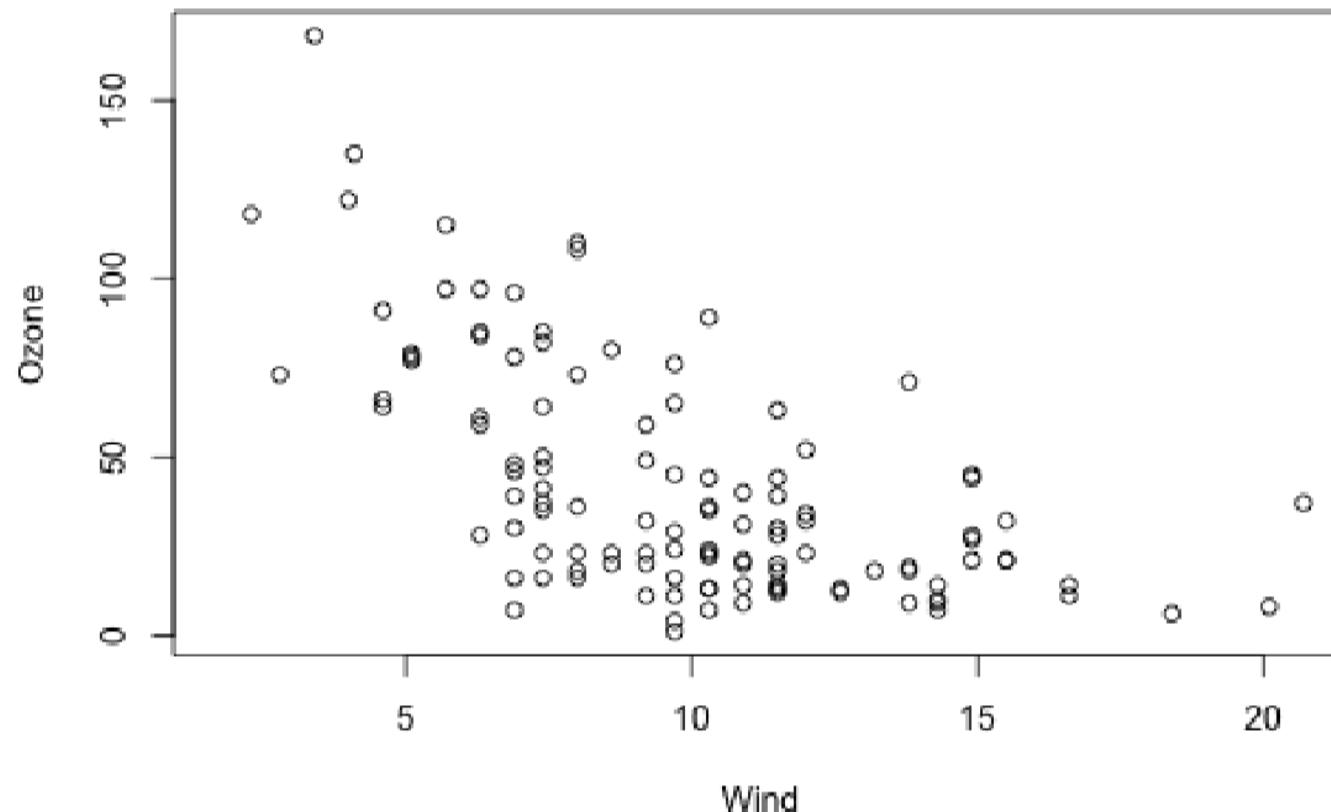
Simple Base Graphics: Histogram

```
library(datasets)  
hist(airquality$Ozone)      ## Draw a new plot
```



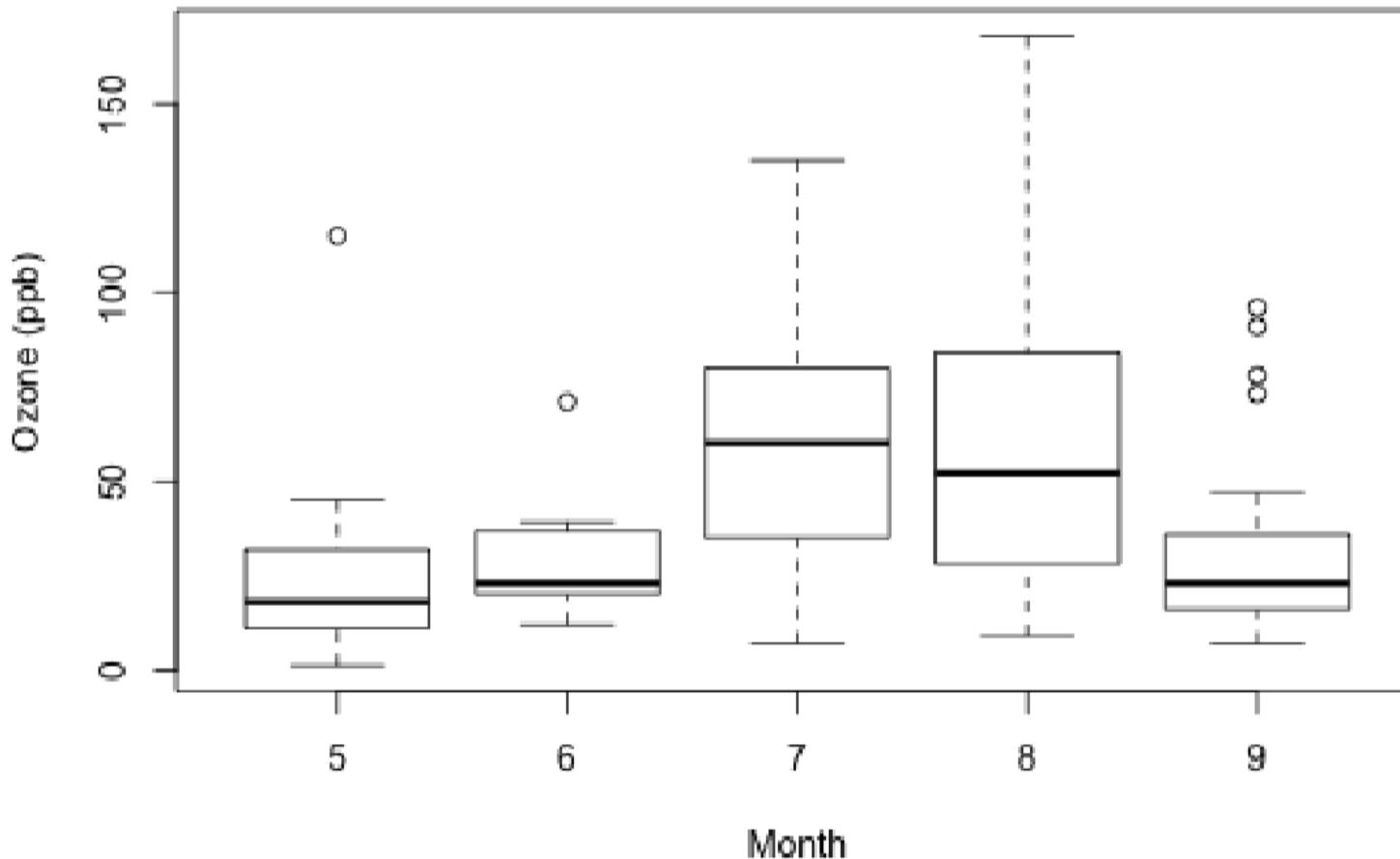
Simple Base Graphics: Scatterplot

```
library(datasets)  
with(airquality, plot(Wind, Ozone))
```



Simple Base Graphics: Boxplot

```
library(datasets)
airquality <- transform(airquality, Month = factor(Month))
boxplot(Ozone ~ Month, airquality, xlab = "Month", ylab = "Ozone (ppb)")
```



Some Important Base Graphics Parameters

Many base plotting functions share a set of parameters. Here are a few key ones:

- `pch`: the plotting symbol (default is open circle)
- `lty`: the line type (default is solid line), can be dashed, dotted, etc.
- `lwd`: the line width, specified as an integer multiple
- `col`: the plotting color, specified as a number, string, or hex code; the `colors()` function gives you a vector of colors by name
- `xlab`: character string for the x-axis label
- `ylab`: character string for the y-axis label

Some Important Base Graphics Parameters

The `par()` function is used to specify *global* graphics parameters that affect all plots in an R session. These parameters can be overridden when specified as arguments to specific plotting functions.

- `las`: the orientation of the axis labels on the plot
- `bg`: the background color
- `mar`: the margin size
- `oma`: the outer margin size (default is 0 for all sides)
- `mfrow`: number of plots per row, column (plots are filled row-wise)
- `mfcol`: number of plots per row, column (plots are filled column-wise)

Some Important Base Graphics Parameters

Default values for global graphics parameters

```
par( "lty" )  
  
## [1] "solid"  
  
par( "col" )  
  
## [1] "black"  
  
par( "pch" )  
  
## [1] 1
```

Some Important Base Graphics Parameters

Default values for global graphics parameters

```
par( "bg" )
```

```
## [1] "transparent"
```

```
par( "mar" )
```

```
## [1] 5.1 4.1 4.1 2.1
```

```
par( "mfrow" )
```

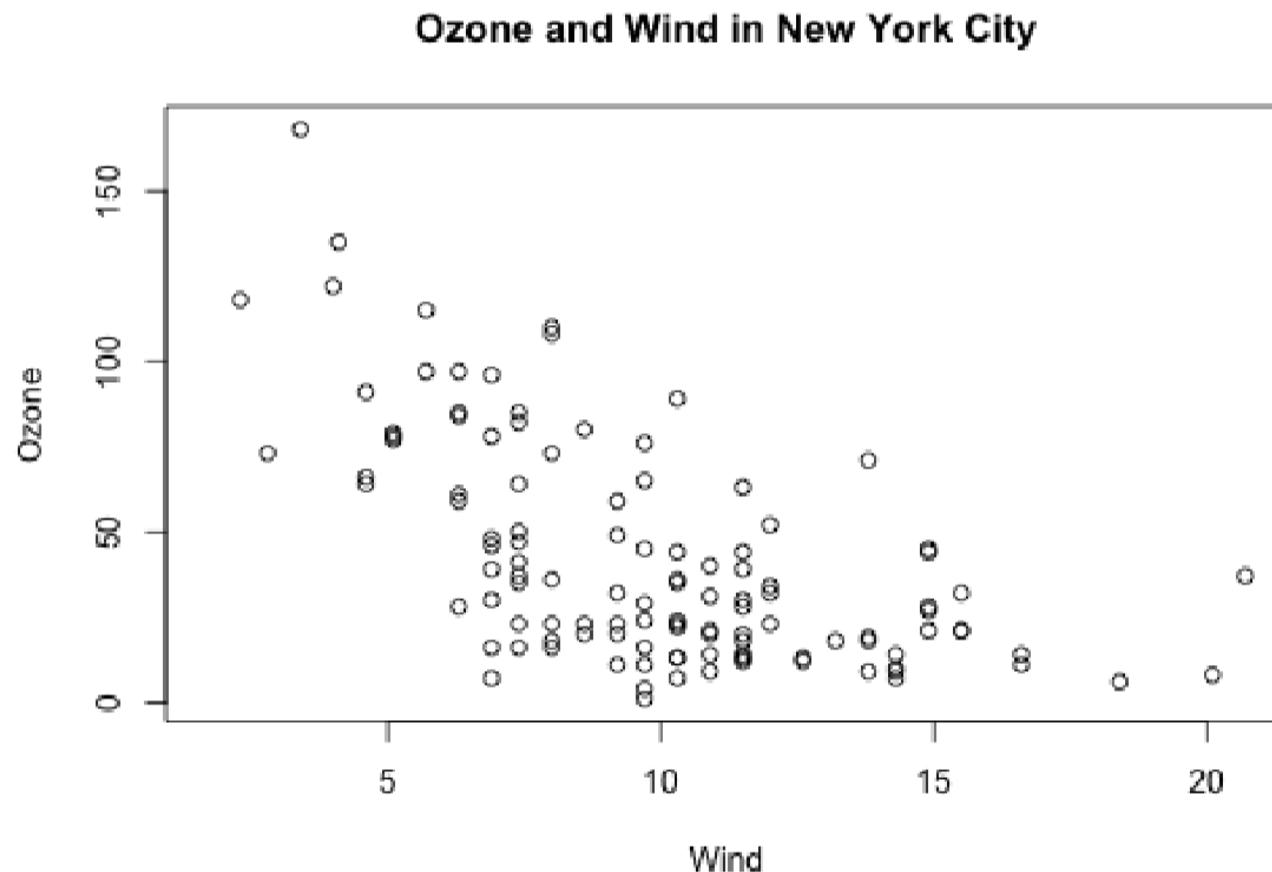
```
## [1] 1 1
```

Base Plotting Functions

- `plot`: make a scatterplot, or other type of plot depending on the class of the object being plotted
- `lines`: add lines to a plot, given a vector `x` values and a corresponding vector of `y` values (or a 2- column matrix); this function just connects the dots
- `points`: add points to a plot
- `text`: add text labels to a plot using specified `x`, `y` coordinates
- `title`: add annotations to `x`, `y` axis labels, title, subtitle, outer margin
- `mtext`: add arbitrary text to the margins (inner or outer) of the plot
- `axis`: adding axis ticks/labels

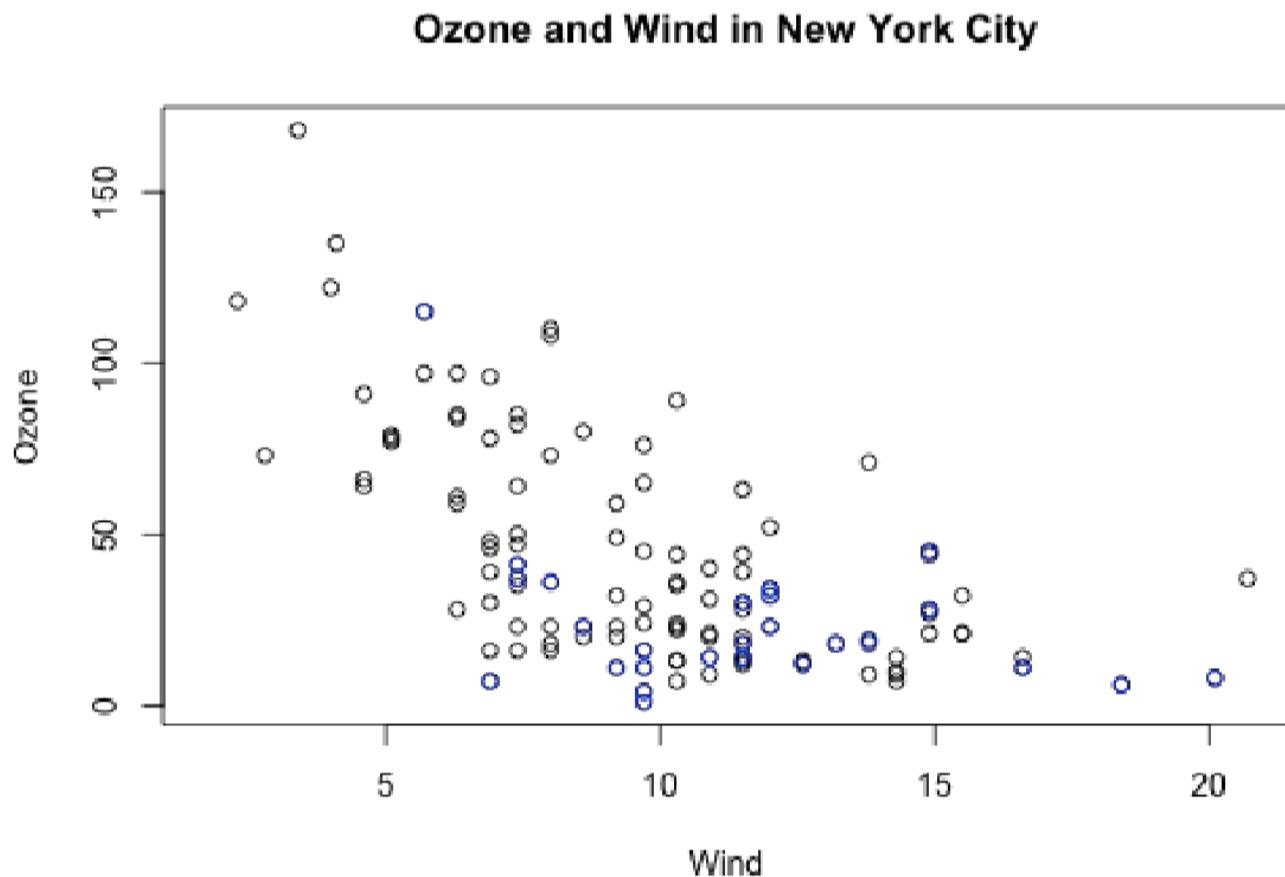
Base Plot with Annotation

```
library(datasets)
with(airquality, plot(Wind, Ozone))
title(main = "Ozone and Wind in New York City") ## Add a title
```



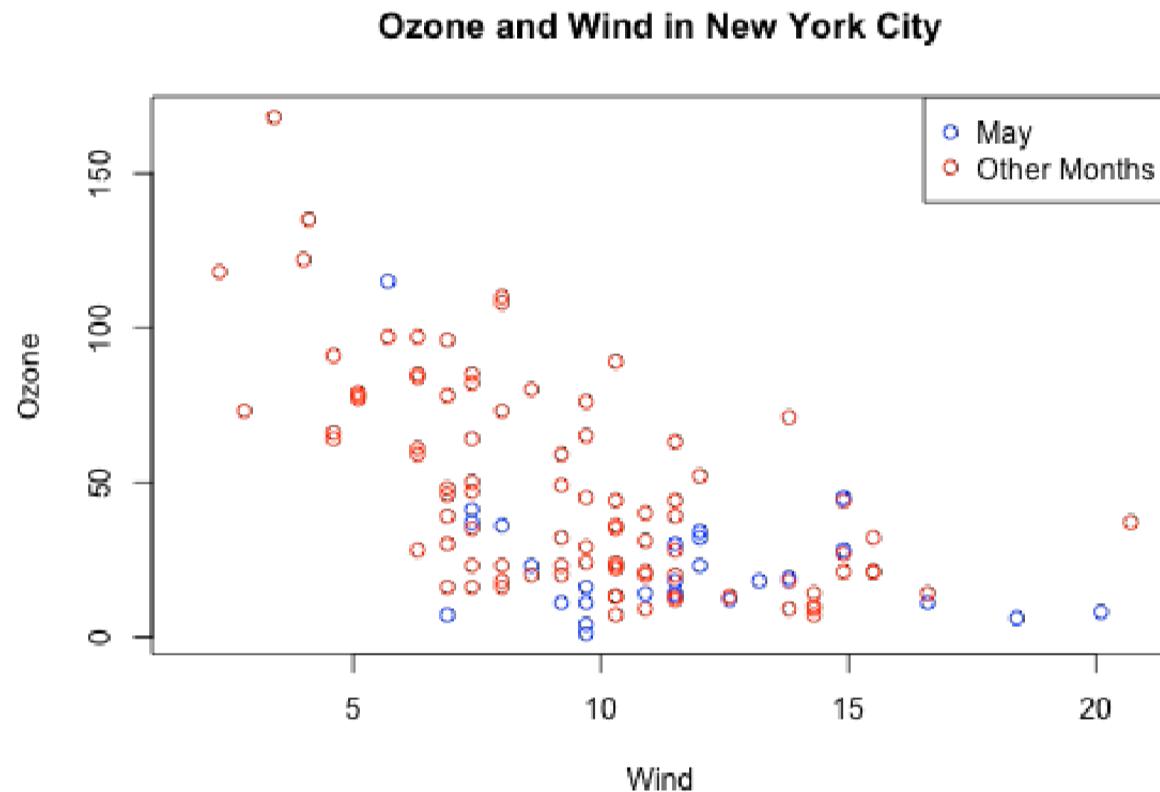
Base Plot with Annotation

```
with(airquality, plot(Wind, Ozone, main = "Ozone and Wind in  
New York City")) with(subset(airquality, Month == 5),  
points(Wind, Ozone, col = "blue"))
```



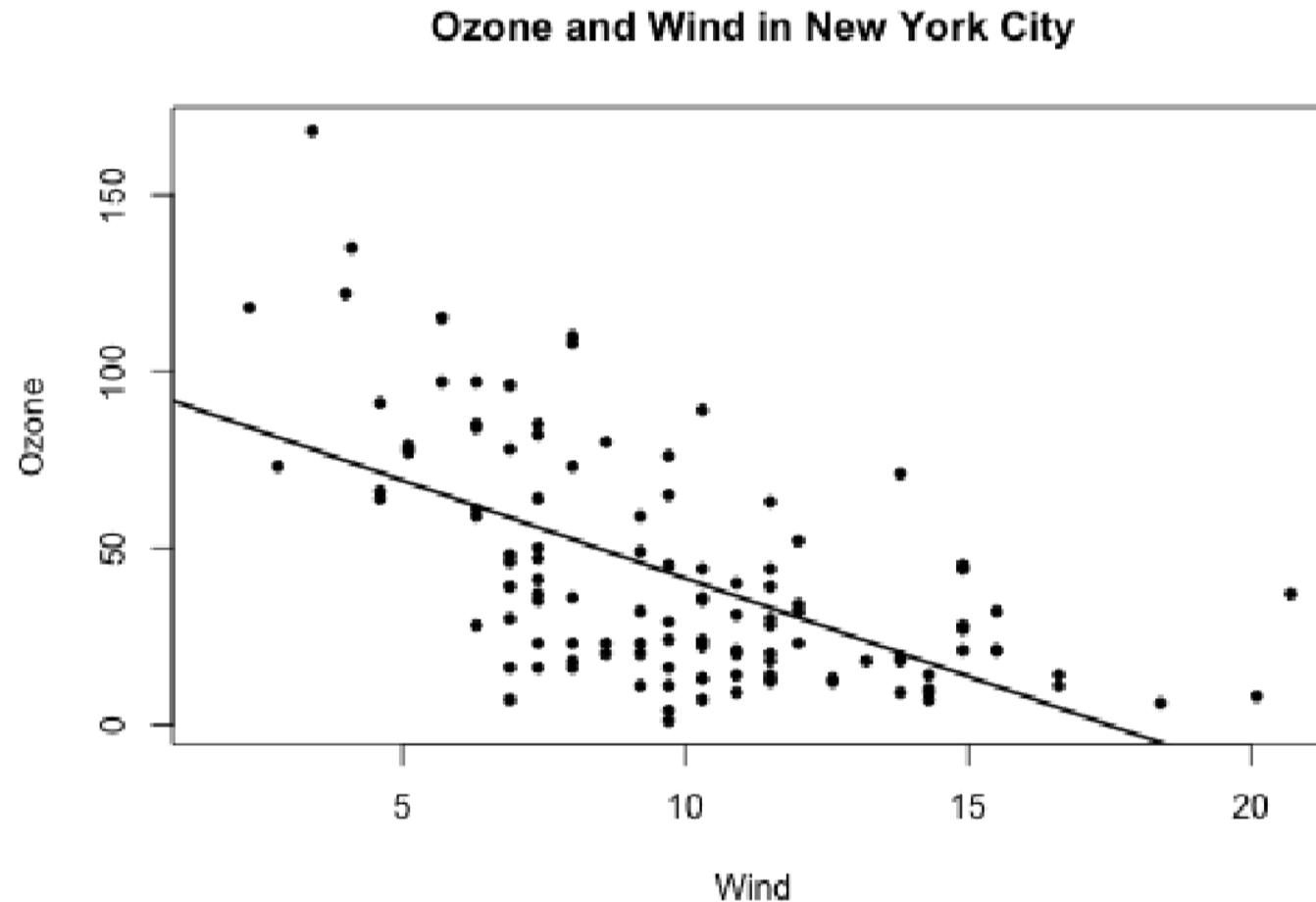
Base Plot with Annotation

```
with(airquality, plot(Wind, Ozone, main = "Ozone and Wind in New York City", type = "n"))
with(subset(airquality, Month == 5), points(Wind, Ozone, col = "blue"))
with(subset(airquality, Month != 5), points(Wind, Ozone, col = "red"))
legend("topright", pch = 1, col = c("blue", "red"), legend = c("May", "Other Months"))
```



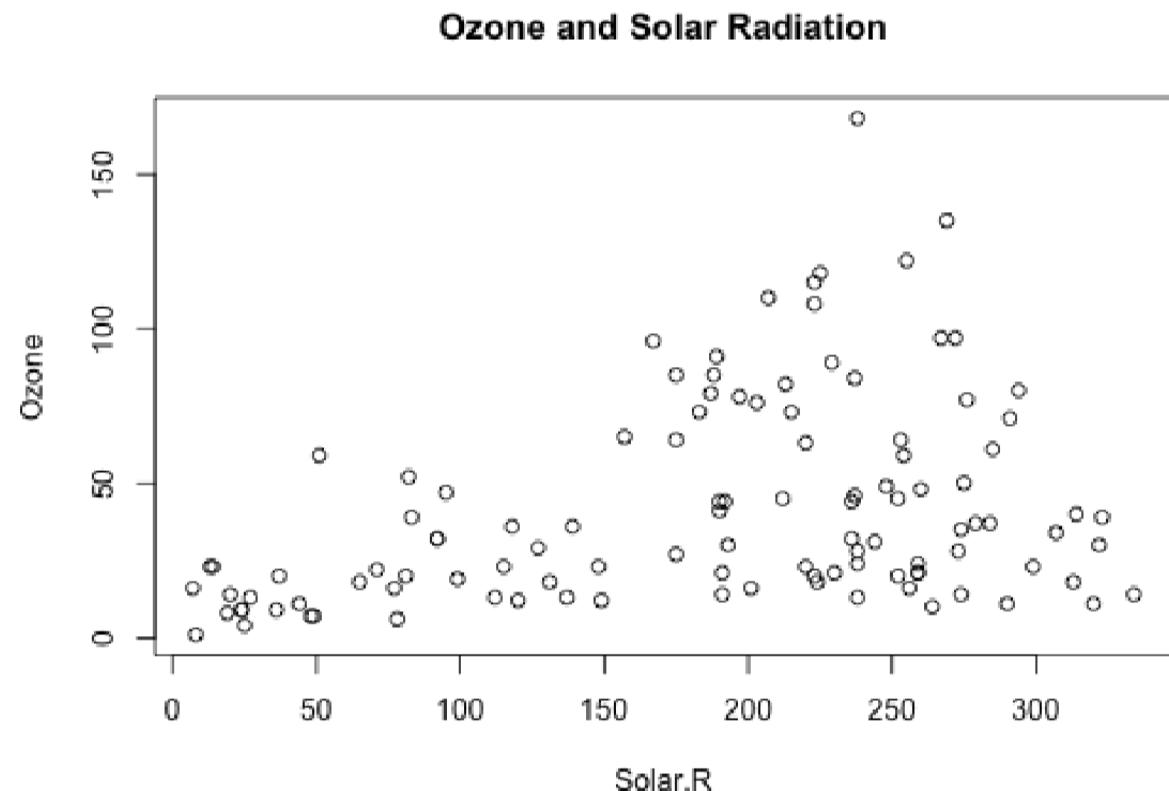
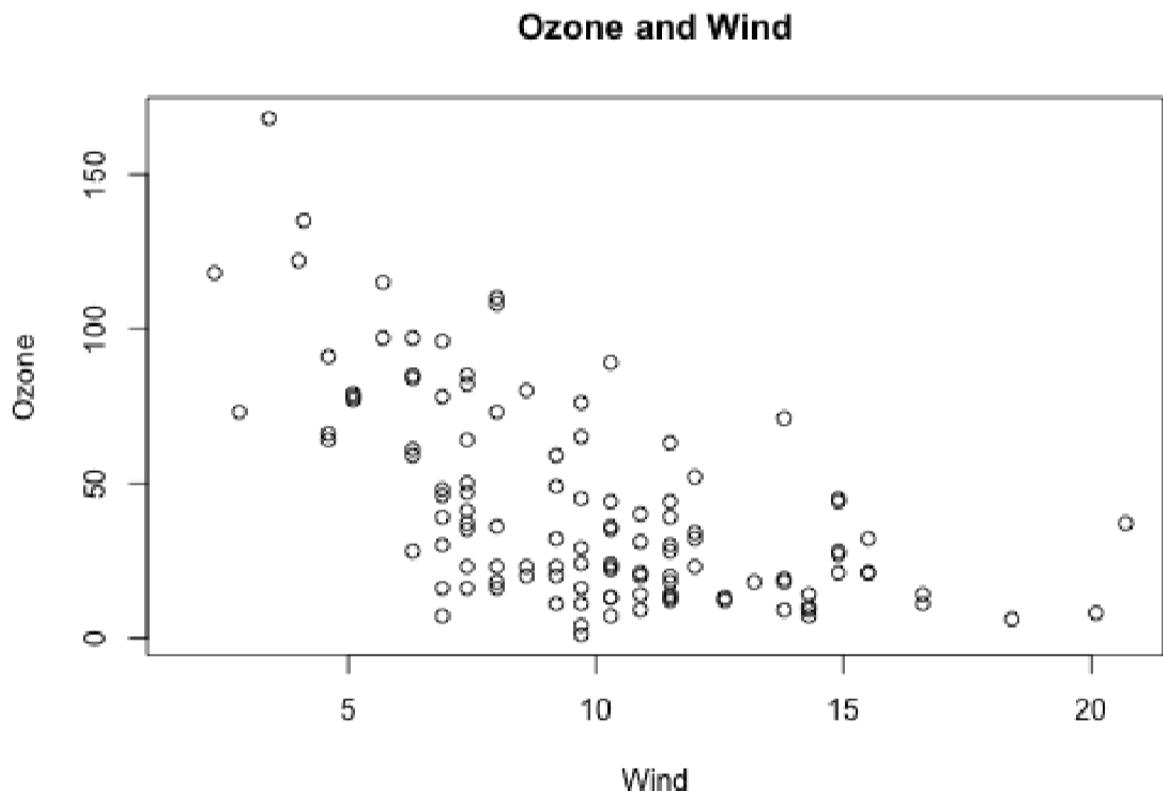
Base Plot with Regression Line

```
with(airquality, plot(Wind, Ozone, main = "Ozone and Wind in New York City", pch = 20))  
model <- lm(Ozone ~ Wind, airquality) abline(model, lwd = 2)
```



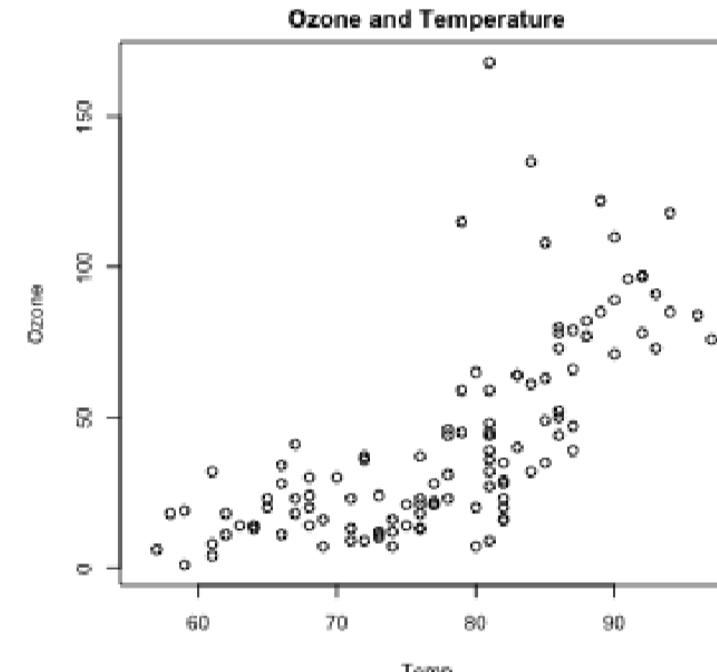
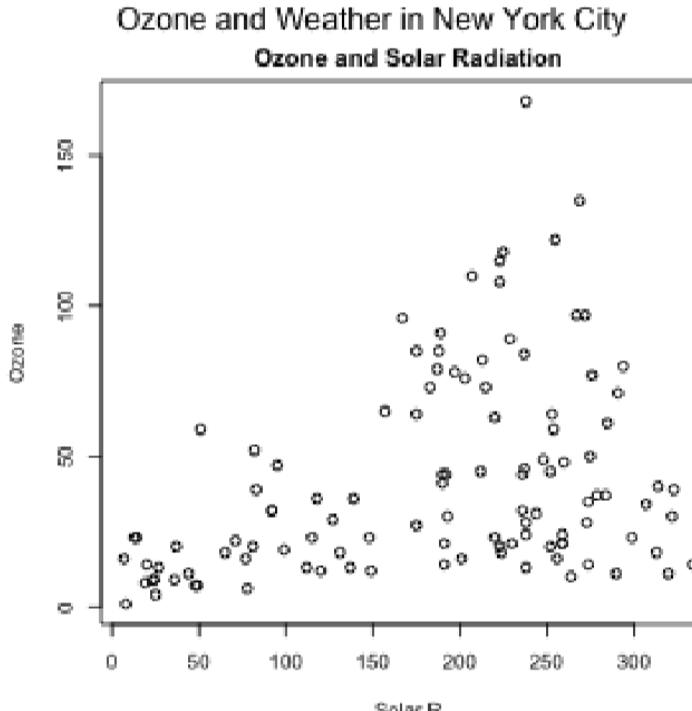
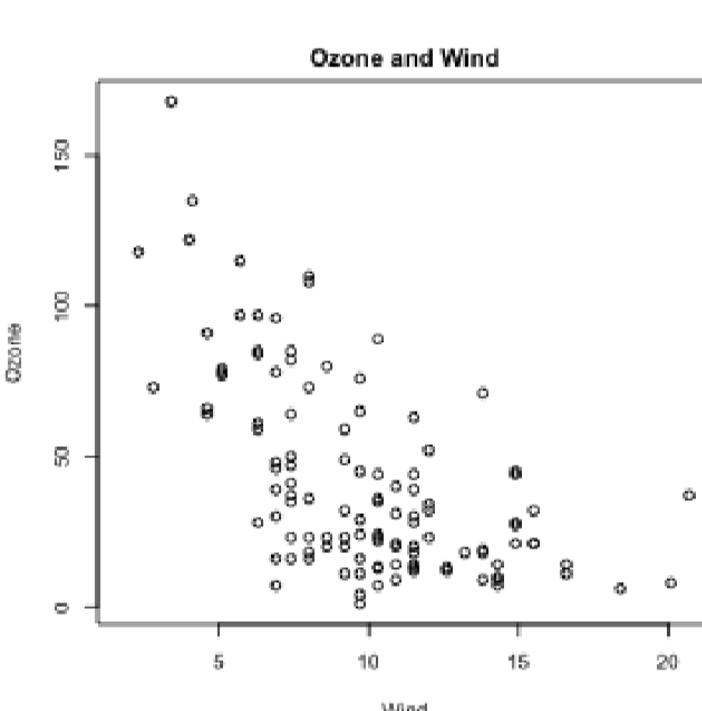
Multiple Base Plots

```
par(mfrow = c(1, 2)) with(airquality, {  
  plot(Wind, Ozone, main = "Ozone and Wind")  plot(Solar.R, Ozone,  
  main = "Ozone and Solar Radiation")  
})
```



Multiple Base Plots

```
par(mfrow = c(1, 3), mar = c(4, 4, 2, 1), oma = c(0, 0, 2, 0)) with(airquality, {  
  plot(Wind, Ozone, main = "Ozone and Wind")  
  plot(Solar.R, Ozone, main = "Ozone and Solar Radiation")  
  plot(Temp, Ozone, main = "Ozone and Temperature")  
  mtext("Ozone and Weather in New York City", outer = TRUE)  
})
```



Base R Plotting Summary

- Plots in the base plotting system are created by calling successive R functions to "build up" a plot
- Plotting occurs in two stages:
 - Creation of a plot
 - Annotation of a plot (adding lines, points, text, legends)
- The base plotting system is very flexible and offers a high degree of control over plotting

Example Data

- There are 104 built-in datasets in R

```
ls("package:datasets")
```

- Let's choose “Theoph”

The screenshot shows the RStudio interface with the 'Console' tab active. The code 'ls("package:datasets")' is entered at the top. The output shows a list of 104 built-in datasets, with 'Theoph' highlighted in yellow.

```
ls("package:datasets")
[1] "ability.cov"          "airmiles"           "AirPassengers"
[4] "airquality"           "anscombe"            "attenu"
[7] "altitude"              "austres"             "beaver1"
[10] "beaver2"               "BJsales"            "BJsales.lead"
[13] "BOD"                  "cars"                "ChickWeight"
[16] "chickwts"              "co2"                 "CO2"
[19] "crimtab"              "discoveries"         "DNase"
[22] "esoph"                 "euro"                "euro.cross"
[25] "eurodist"              "EuStockMarkets"     "faithful"
[28] "fdeaths"               "Formaldehyde"       "freeny"
[31] "freeny.x"              "freeny.y"            "HairEyeColor"
[34] "Harman23.cor"          "Harman74.cor"        "Indometh"
[37] "infert"                 "InsectSprays"        "iris"
[40] "iris3"                  "islands"             "JohnsonJohnson"
[43] "LakeHuron"              "ldeaths"             "lh"
[46] "LifeCycleSavings"       "Loblolly"             "longley"
[49] "lynx"                   "mdeaths"             "morley"
[52] "mtcars"                 "nhtemp"              "Nile"
[55] "nottem"                 "npk"                 "occupationalStatus"
[58] "Orange"                 "OrchardSprays"       "PlantGrowth"
[61] "precip"                 "presidents"          "pressure"
[64] "Puromycin"              "quakes"              "randu"
[67] "rivers"                 "rock"                "Seatbelts"
[70] "sleep"                  "stack.loss"           "stack.x"
[73] "stackloss"              "state.abb"            "state.area"
[76] "state.center"           "state.division"       "state.name"
[79] "state.region"            "state.x77"            "sunspot.month"
[82] "sunspot.year"           "sunspots"             "swiss"
[85] "Theoph"                  "Titanic"              "ToothGrowth"
[88] "treering"                "trees"                "UCBAdmissions"
[91] "UKDriverDeaths"          "UKgas"                "USAccDeaths"
[94] "USArests"                "UScitiesD"            "USJudgeRatings"
[97] "USPersonalExpenditure"   "uspop"                "VADeaths"
[100] "volcano"                 "warpbreaks"           "women"
[103] "WorldPhones"             "WWWusage"
```

Theoph Dataset Base Plotting Example

- Theoph is one of many example datasets in R
- Basic R maneuvers
- Names and number of rows/columns
- Attach/detach dataframe
- Subsetting
- Descriptive Statistics
- Base plotting recap
- Intro to complex plotting with ggplot2

Basic R Maneuvers

- View the first and last several rows of “Theoph” dataset
- **12 subjects**, consisting of **132 data points** with theophylline concentrations up to **24 hr post dose**

```
10
11 ##### Output the first few rows of a dataframe-----
12 head(Theoph)
13
14 ##### Output the last few rows of a dataframe-----
15 tail(Theoph)
```

```
> head(Theoph)
Grouped Data: conc ~ Time | Subject
  Subject   Wt Dose Time conc
  1       1 79.6 4.02 0.00 0.74
  2       1 79.6 4.02 0.25 2.84
  3       1 79.6 4.02 0.57 6.57
  4       1 79.6 4.02 1.12 10.50
  5       1 79.6 4.02 2.02 9.66
  6       1 79.6 4.02 3.82 8.58
> tail(Theoph)
Grouped Data: conc ~ Time | Subject
  Subject   Wt Dose Time conc
  127      12 60.5 5.3  3.52 9.75
  128      12 60.5 5.3  5.07 8.57
  129      12 60.5 5.3  7.07 6.59
  130      12 60.5 5.3  9.03 6.11
  131      12 60.5 5.3 12.05 4.57
  132      12 60.5 5.3 24.15 1.17
```

- Can assess # of rows, columns

```
#### Check the number of rows-----
nrow(Theoph)

#### Check the number of columns-----
ncol(Theoph)

> nrow(Theoph)
[1] 132
> ncol(Theoph)
[1] 5
```

- View data from a single row, column, cell

```
#### To look at same place in the middle of dataframe, use [row,column]-----
Theoph[51,]
Theoph[,4]
```

```
#### To look at a particular element in dataset-----
Theoph[51,2]
```

```
> Theoph[51,]
Grouped Data: conc ~ Time | subject
  Subject   Wt Dose Time conc
51      5 54.6 5.86 5.02 7.56
> Theoph[,4]
[1]  0.00  0.25  0.57  1.12  2.02  3.82  5.10  7.03  9.05 12.12 24.37  0.00  0.27  0.52  1.00  1.92  3.50
[18] 5.02  7.03  9.00 12.00 24.30  0.00  0.27  0.58  1.02  2.02  3.62  5.08  7.07  9.00 12.15 24.17  0.00
[35] 0.35  0.60  1.07  2.13  3.50  5.02  7.02  9.02 11.98 24.65  0.00  0.30  0.52  1.00  2.02  3.50  5.02
[52] 7.02  9.10 12.00 24.35  0.00  0.27  0.58  1.15  2.03  3.57  5.00  7.00  9.22 12.10 23.85  0.00  0.25
[69] 0.50  1.02  2.02  3.48  5.00  6.98  9.00 12.05 24.22  0.00  0.25  0.52  0.98  2.02  3.53  5.05  7.15
[86] 9.07 12.10 24.12  0.00  0.30  0.63  1.05  2.02  3.53  5.02  7.17  8.80 11.60 24.43  0.00  0.37  0.77
[103] 1.02  2.05  3.55  5.05  7.08  9.38 12.10 23.70  0.00  0.25  0.50  0.98  1.98  3.60  5.02  7.03  9.03
[120] 12.12 24.08  0.00  0.25  0.50  1.00  2.00  3.52  5.07  7.07  9.03 12.05 24.15
> Theoph[51,2]
[1] 54.6
```

- View column names

```
##### Check column names for a dataframe-----
names(Theoph)

->
> names(Theoph)
[1] "subject"    "wt"       "Dose"     "Time"     "conc"
```

- Can attach, detach dataframes
 - Once attached, don't need to reference what dataframe you're working with
 - Need to detach before using another dataframe

```
#### attach.dataframe, but be careful to detach before using another dataset-----  
attach(Theoph)  
detach(Theoph)
```

- Can also reference dataframe each line without attaching

```
#### to look for a specific column in a dataframe without attaching dataframe----  
Theoph$Time
```

- Let's say you want to know how many time measurements were made within the first 3 hrs post dose

```
#### How many Time measurements are less than 3 hrs post dose?-----
```

```
Theoph[Theoph$Time<3, ]
```

```
Console C:/Users/peerc/Desktop/NCI Projects/Classes/FAES/BioTech84_ExpRespUsingR_March2017/R Training/ 
71    7 64.6 4.95 2.02  6.58
78    8 70.5 4.53 0.00  0.00
79    8 70.5 4.53 0.25  3.05
80    8 70.5 4.53 0.52  3.05
81    8 70.5 4.53 0.98  7.31
82    8 70.5 4.53 2.02  7.56
89    9 86.4 3.10 0.00  0.00
90    9 86.4 3.10 0.30  7.37
91    9 86.4 3.10 0.63  9.03
92    9 86.4 3.10 1.05  7.14
93    9 86.4 3.10 2.02  6.33
100   10 58.2 5.50 0.00  0.24
101   10 58.2 5.50 0.37  2.89
102   10 58.2 5.50 0.77  5.22
103   10 58.2 5.50 1.02  6.41
104   10 58.2 5.50 2.05  7.83
111   11 65.0 4.92 0.00  0.00
112   11 65.0 4.92 0.25  4.86
113   11 65.0 4.92 0.50  7.24
114   11 65.0 4.92 0.98  8.00
115   11 65.0 4.92 1.98  6.81
122   12 60.5 5.30 0.00  0.00
123   12 60.5 5.30 0.25  1.25
124   12 60.5 5.30 0.50  3.96
125   12 60.5 5.30 1.00  7.82
126   12 60.5 5.30 2.00  9.72
> |
```

- Not practical, so let's rephrase the question...asking for the number of rows

```
### How many rows?--
```

```
nrow(Theoph[Theoph$Time<3, ])
```

```
> nrow(Theoph[Theoph$Time<3, ])
[1] 60
> |
```

- How about a more relevant question

```
#### How many Time measurements are at pre-dose (Time zero)?----  
Theoph[Theoph$Time==0,]  
nrow(Theoph[Theoph$Time==0,])  
  
> Theoph[Theoph$Time==0,]  
Grouped Data: conc ~ Time | Subject  
  Subject   Wt Dose Time conc  
1       1 79.6 4.02    0 0.74  
12      2 72.4 4.40    0 0.00  
23      3 70.5 4.53    0 0.00  
34      4 72.7 4.40    0 0.00  
45      5 54.6 5.86    0 0.00  
56      6 80.0 4.00    0 0.00  
67      7 64.6 4.95    0 0.15  
78      8 70.5 4.53    0 0.00  
89      9 86.4 3.10    0 0.00  
100     10 58.2 5.50   0 0.24  
111     11 65.0 4.92   0 0.00  
122     12 60.5 5.30   0 0.00  
> nrow(Theoph[Theoph$Time==0,])  
[1] 12  
> |
```

- All 12 subjects do. But how many subjects had measurable theophylline concentrations at time zero?

```
### Instead of all data, ask how many rows of data have conc data above 0 at time zero----  
Theoph[Theoph$Time==0 & Theoph$conc >0,]  
nrow(Theoph[Theoph$Time==0 & Theoph$conc >0,])  
  
> Theoph[Theoph$Time==0 & Theoph$conc >0,]  
Grouped Data: conc ~ Time | Subject  
  Subject   Wt Dose Time conc  
1       1 79.6 4.02    0 0.74  
67      7 64.6 4.95    0 0.15  
100     10 58.2 5.50   0 0.24  
> nrow(Theoph[Theoph$Time==0 & Theoph$conc >0,])  
[1] 3
```

Other relevant information...

```
#### Try to get weights of individuals between 40 and 60-----  
names(Theoph)  
Theoph[Theoph$wt > 40 & Theoph$wt < 60,]
```

```
[1] "Subject" "wt"      "Dose"    "Time"    "conc"  
> Theoph[Theoph$wt > 40 & Theoph$wt < 60,]  
Grouped Data: conc ~ Time | Subject  
  Subject   wt Dose  Time conc  
45       5 54.6 5.86 0.00 0.00  
46       5 54.6 5.86 0.30 2.02  
47       5 54.6 5.86 0.52 5.63  
48       5 54.6 5.86 1.00 11.40  
49       5 54.6 5.86 2.02 9.33  
50       5 54.6 5.86 3.50 8.74  
51       5 54.6 5.86 5.02 7.56  
52       5 54.6 5.86 7.02 7.09  
53       5 54.6 5.86 9.10 5.90  
54       5 54.6 5.86 12.00 4.37  
55       5 54.6 5.86 24.35 1.57  
100      10 58.2 5.50 0.00 0.24  
101      10 58.2 5.50 0.37 2.89  
102      10 58.2 5.50 0.77 5.22  
103      10 58.2 5.50 1.02 6.41  
104      10 58.2 5.50 2.05 7.83  
105      10 58.2 5.50 3.55 10.21  
106      10 58.2 5.50 5.05 9.18  
107      10 58.2 5.50 7.08 8.02  
108      10 58.2 5.50 9.38 7.14  
109      10 58.2 5.50 12.10 5.68  
110      10 58.2 5.50 23.70 2.42  
> |
```

```
#### Calculating some summary statistics-----  
mean(Theoph$wt)  
median(Theoph$wt)  
sd(Theoph$wt)  
var (Theoph$wt)  
  
mean(Theoph$wt)  
1] 69.58333  
median(Theoph$wt)  
1] 70.5  
sd(Theoph$wt)  
1] 9.133181  
var (Theoph$wt)  
1] 83.41499
```

Loading Dataset

- R works best with csv files
- If have a dataset in Excel®, make sure you save as a .csv file
- In RStudio, click “Session”, “Set Working Directory”, then browse for the folder on your computer that contains that dataset csv file
- Once WD set, no longer need to type in full computer file directory
 - Example: C/users/doej/desktop/projectX/datafile1
- Will only need to type “datafile1<-read.csv(“datafile1.csv”, header=TRUE)
 - Named “datafile1” in the R workspace, which can be saved
 - Everything you type in an R script can be saved as well

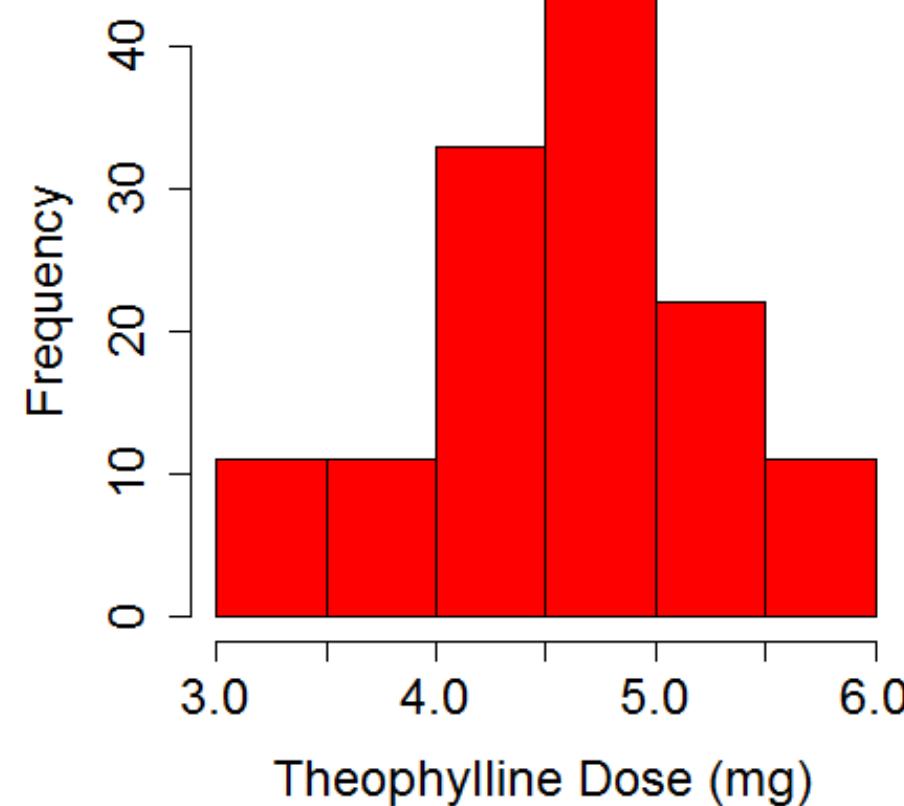
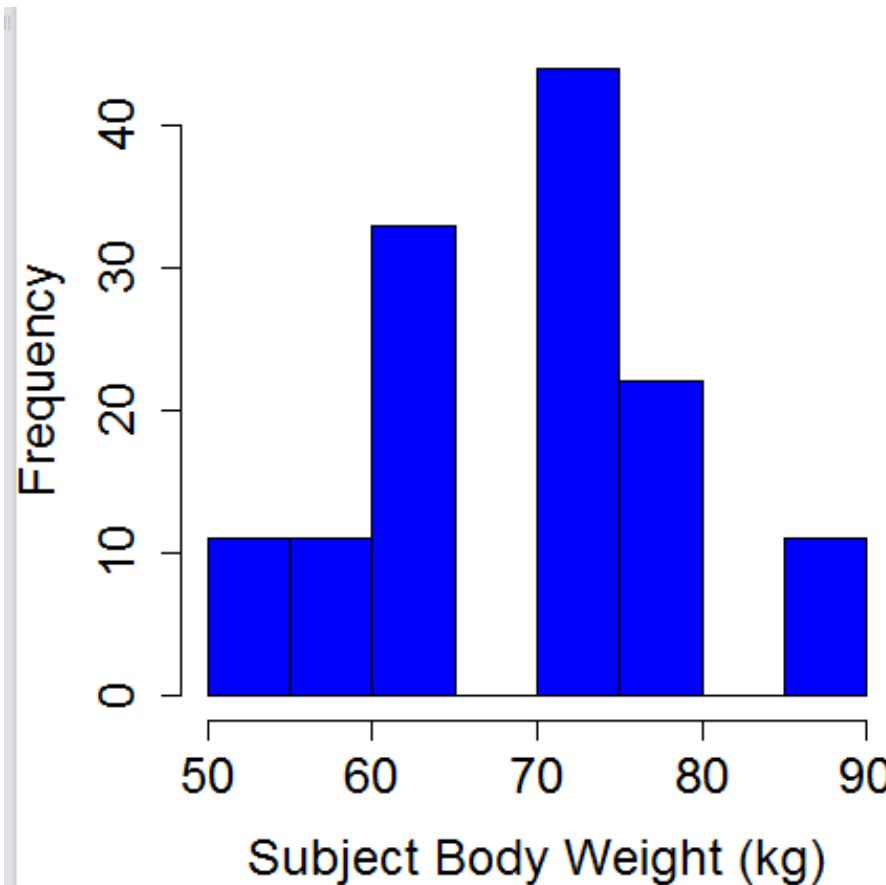
Installing Packages

- Two ways
1. Type *install.packages("name of pkg")*
 2. Click “Packages” tab on lower right side, then “Install”. Search for package name

Let's Plot!

- Histogram of patient weights and doses, side by side

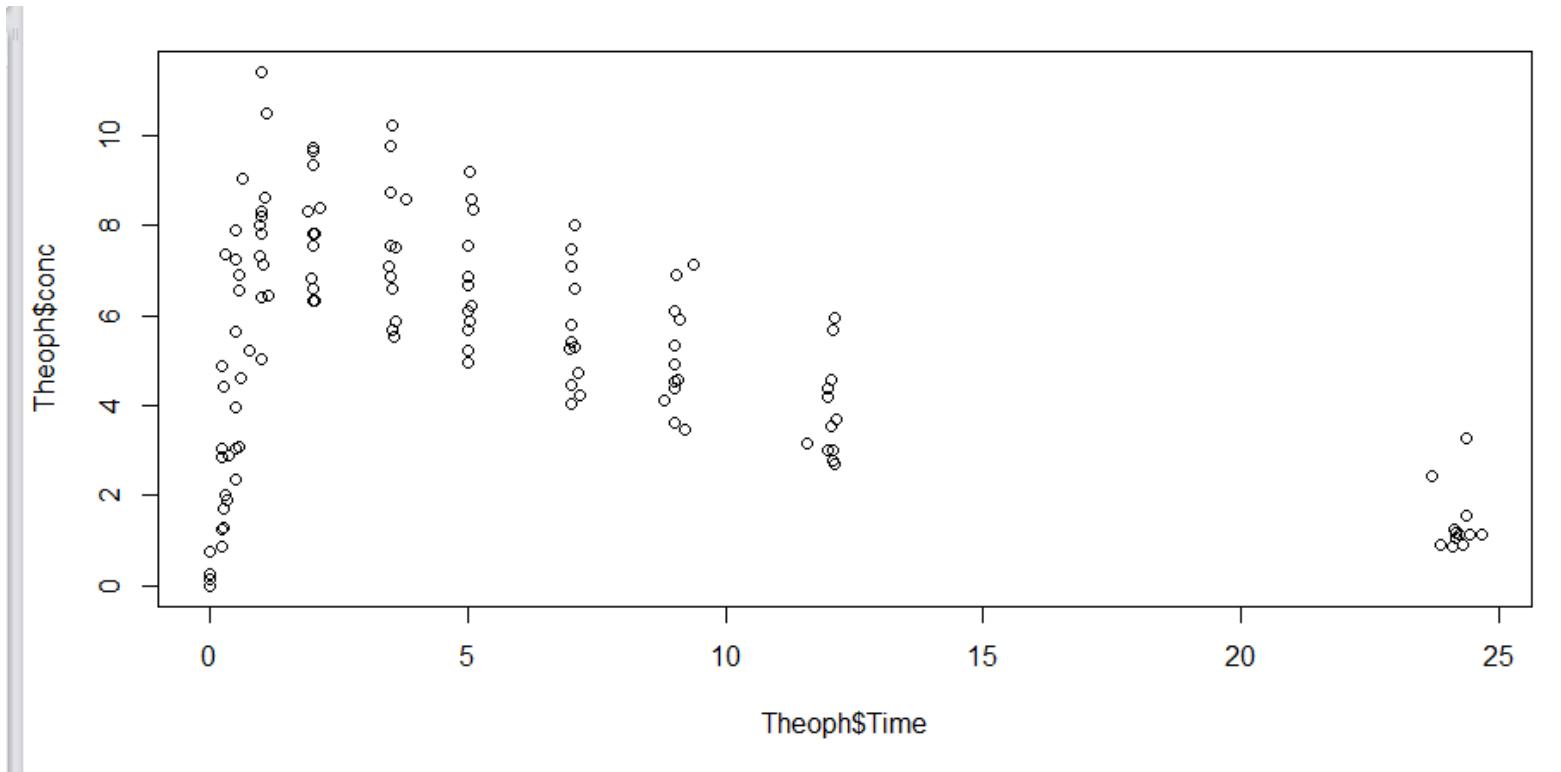
```
### 1 row, 2 columns of plots ###
par(mfrow=c(1,2))
hist(Theoph$wt, cex=1.5, xlab="Subject Body weight (kg)", cex.lab=1.5, cex.axis=1.5, col="blue")
hist(Theoph$Dose, cex=1.5, xlab="Theophylline Dose (mg)", cex.lab=1.5, cex.axis=1.5, col="red")
```



Let's Plot!

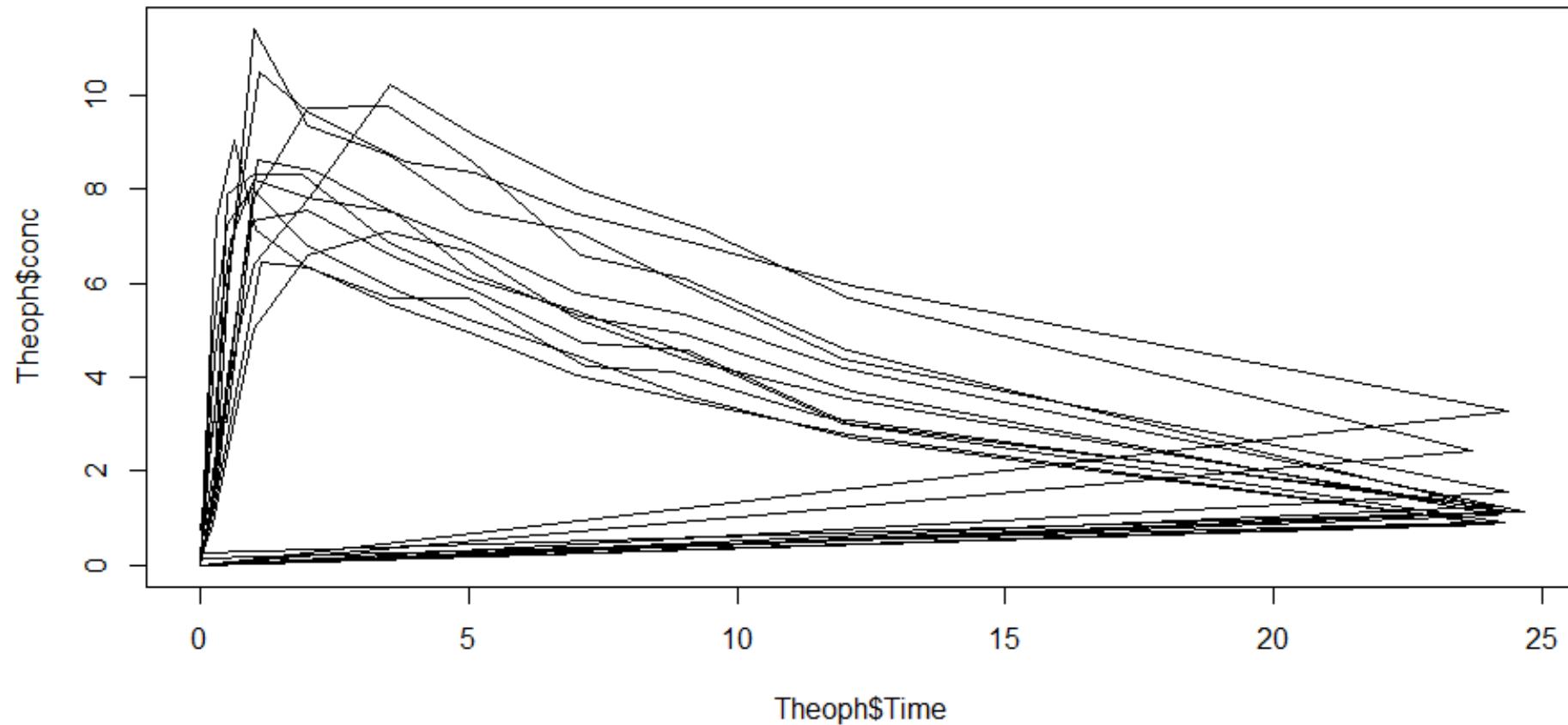
- A simple scatterplot

```
##### Basic plotting function-----
plot(Theoph$Time, Theoph$conc, type="p")
```



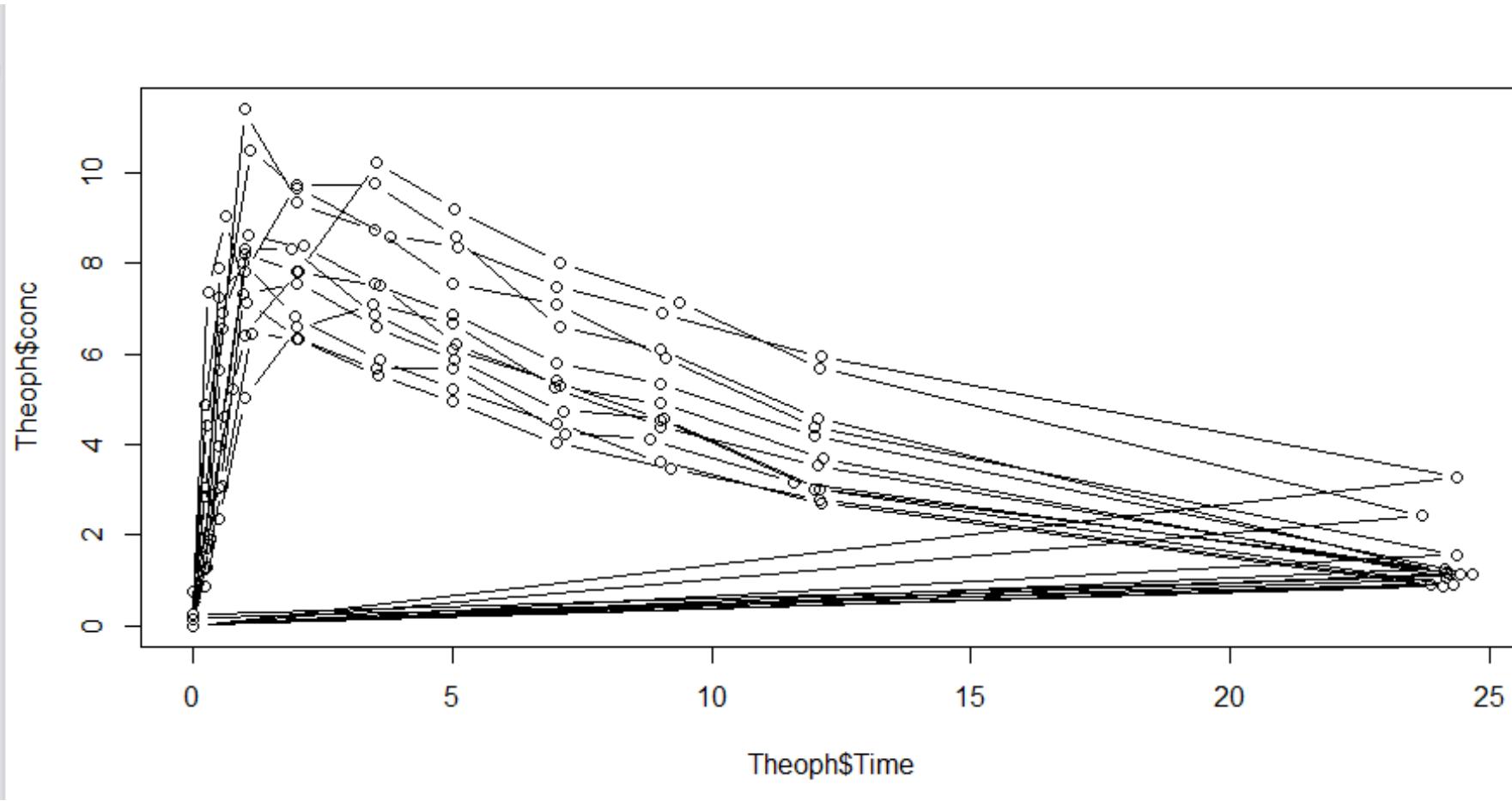
- A line plot

```
plot(Theoph$Time, Theoph$conc, type="l")
```

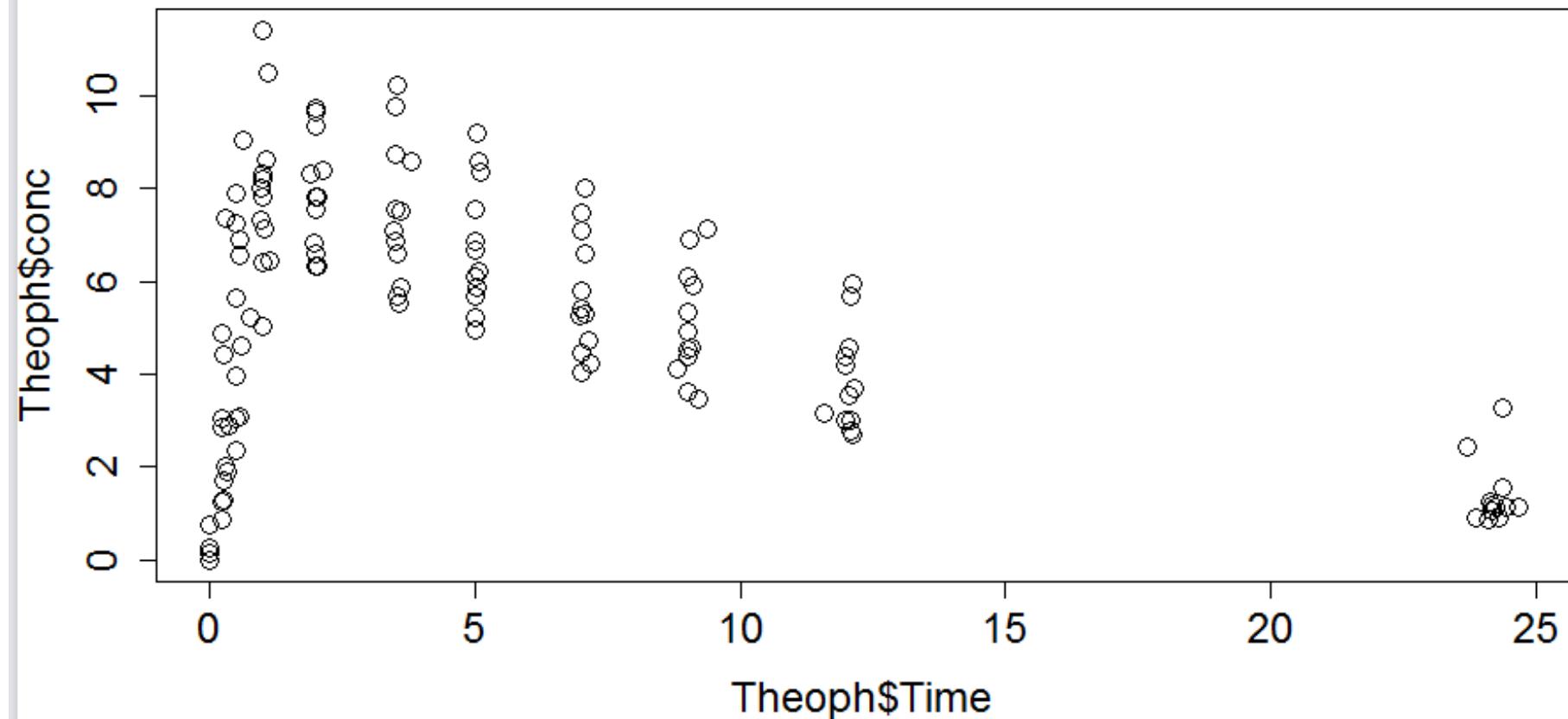


- Both points and lines...

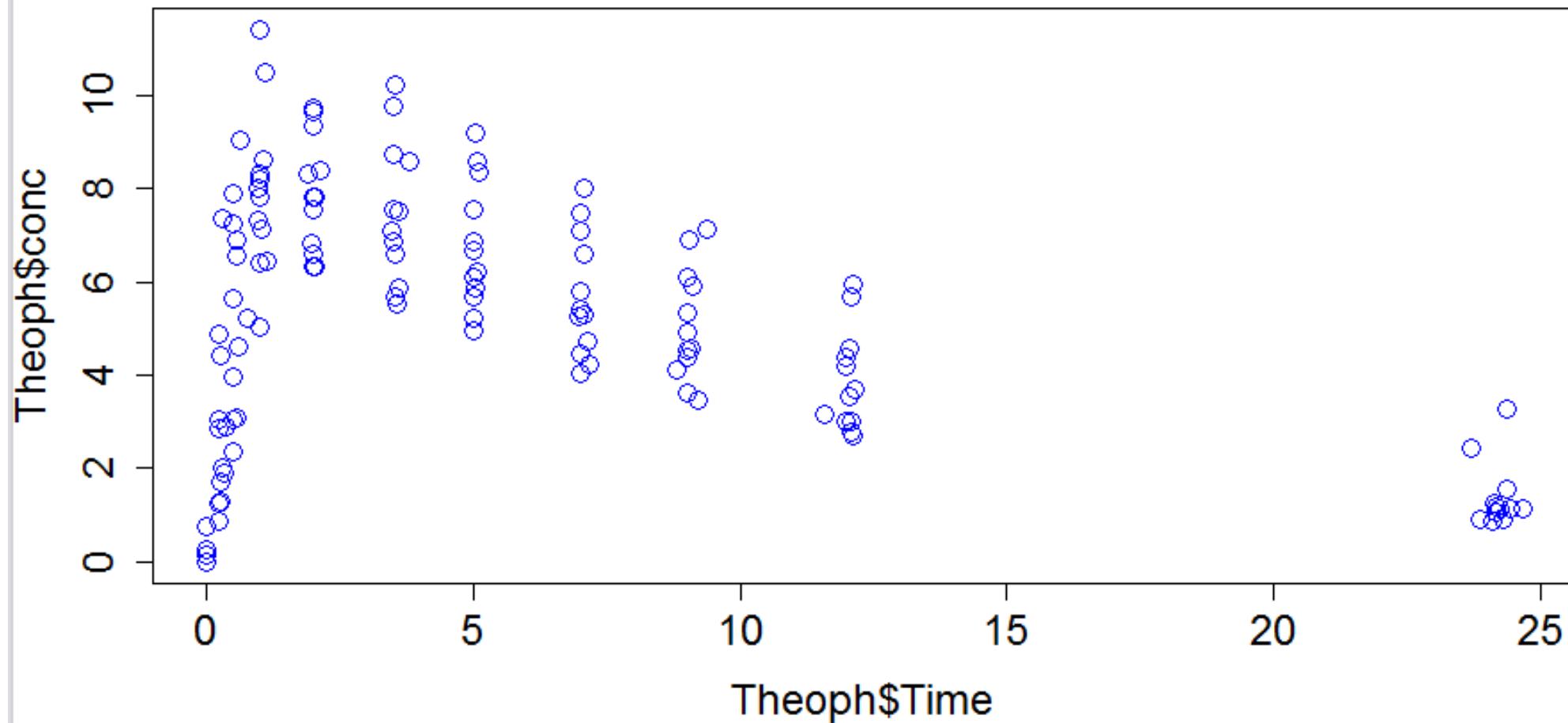
```
plot(Theoph$Time, Theoph$conc, type="b")
```



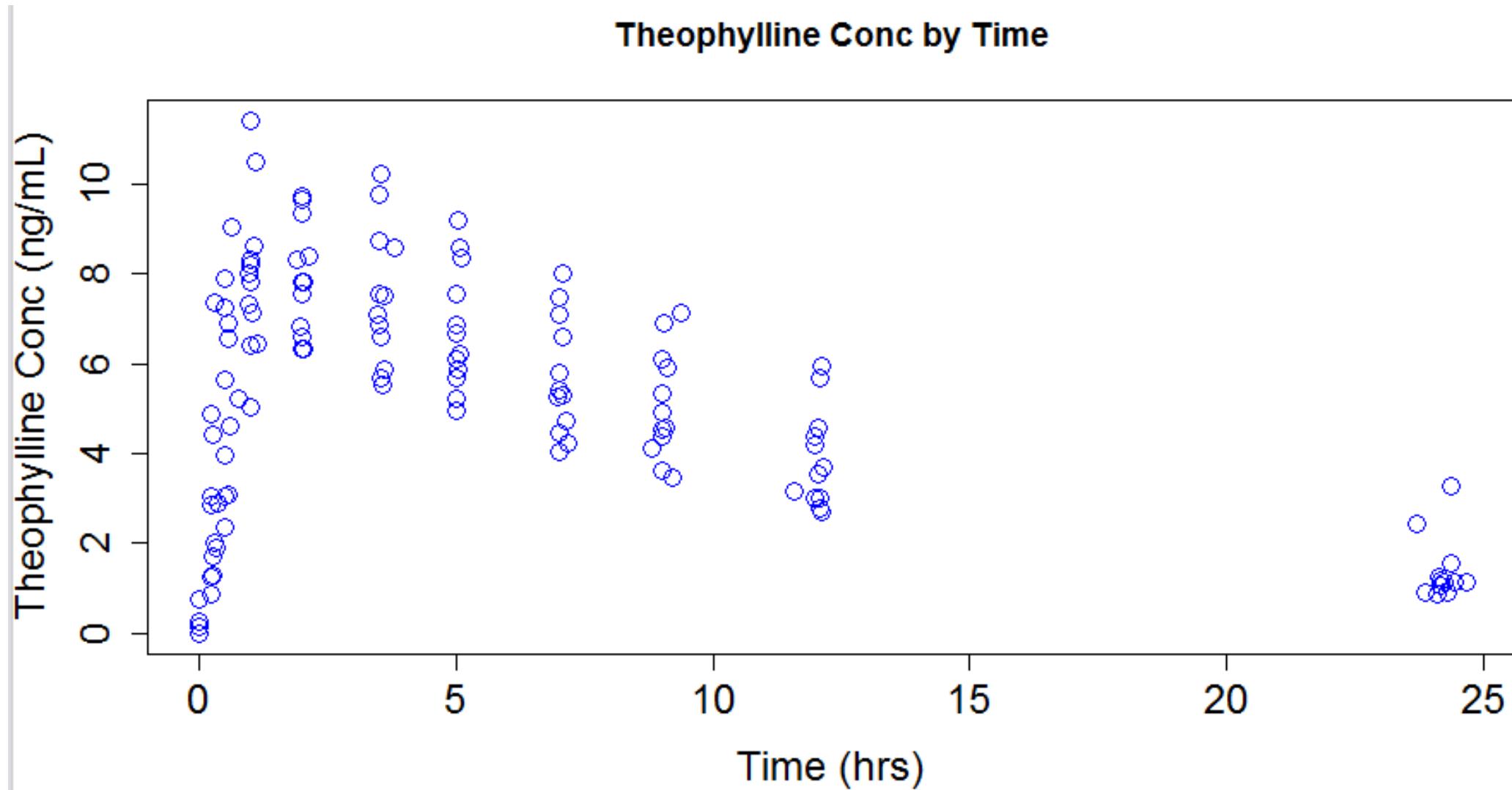
```
### Change size of data points and axes (cex default size =1)###  
plot(Theoph$Time, Theoph$conc, type="p", cex=1.5,  
      cex.lab=1.5, cex.axis=1.5)
```



```
### Add color to data points ###
plot(Theoph$Time, Theoph$conc, type="p", cex=1.5,
      cex.lab=1.5, cex.axis=1.5, col="blue")
```



```
### Add axis labels ###
plot(Theoph$Time, Theoph$conc, type="p", cex=1.5,
      cex.lab=1.5, cex.axis=1.5, col="blue",
      xlab="Time (hrs)", ylab="Theophylline Conc (ng/mL)", main="Theophylline Conc by Time")
```



R plotting with ggplot2

- What is ggplot2?
- Grammar of Graphics
- qplot()
 - Scatter plot
 - Aesthetics
 - Geometric shapes
 - Histogram
 - Facets
 - Density Smooth
- ggplot()
 - geom_point()
 - geom_smooth()
 - Facets
 - Annotation
 - Aesthetics
 - Labels
 - Themes
 - Axis limits

What is ggplot2?

- An implementation of the *Grammar of Graphics* by Leland Wilkinson
- Written by Hadley Wickham (while he was a graduate student at Iowa State)
- A separate graphics system for R (different from **base**)
- Available from CRAN via `install.packages()`
- Web site: <http://ggplot2.org> (better documentation)

What is ggplot2?

- Grammar of graphics represents and abstraction of graphics ideas/objects
- Think “verb”, “noun”, “adjective” for graphics
- Allows for a “theory” of graphics on which to build new graphics and graphics objects
- “Shorten the distance from mind to page”

Grammar of Graphics

“In brief, the grammar tells us that a statistical graphic is a **mapping** from data to **aesthetic** attributes (colour, shape, size) of **geometric** objects (points, lines, bars). The plot may also contain statistical transformations of the data and is drawn on a specific coordinate system”

from *ggplot2* book

The Basics: `qplot()`

- Works much like the `plot` function in base graphics system
- Looks for data in a data frame, similar to `lattice`, or in the parent environment
- Plots are made up of *aesthetics* (size, shape, color) and *geoms* (points, lines)

The Basics: `qplot()`

- Factors are important for indicating subsets of the data (if they are to have different properties); they should be **labeled**
- The `qplot()` hides what goes on underneath, which is okay for most operations
- `ggplot()` is the core function and very flexible for doing things `qplot()` cannot do

Example Dataset

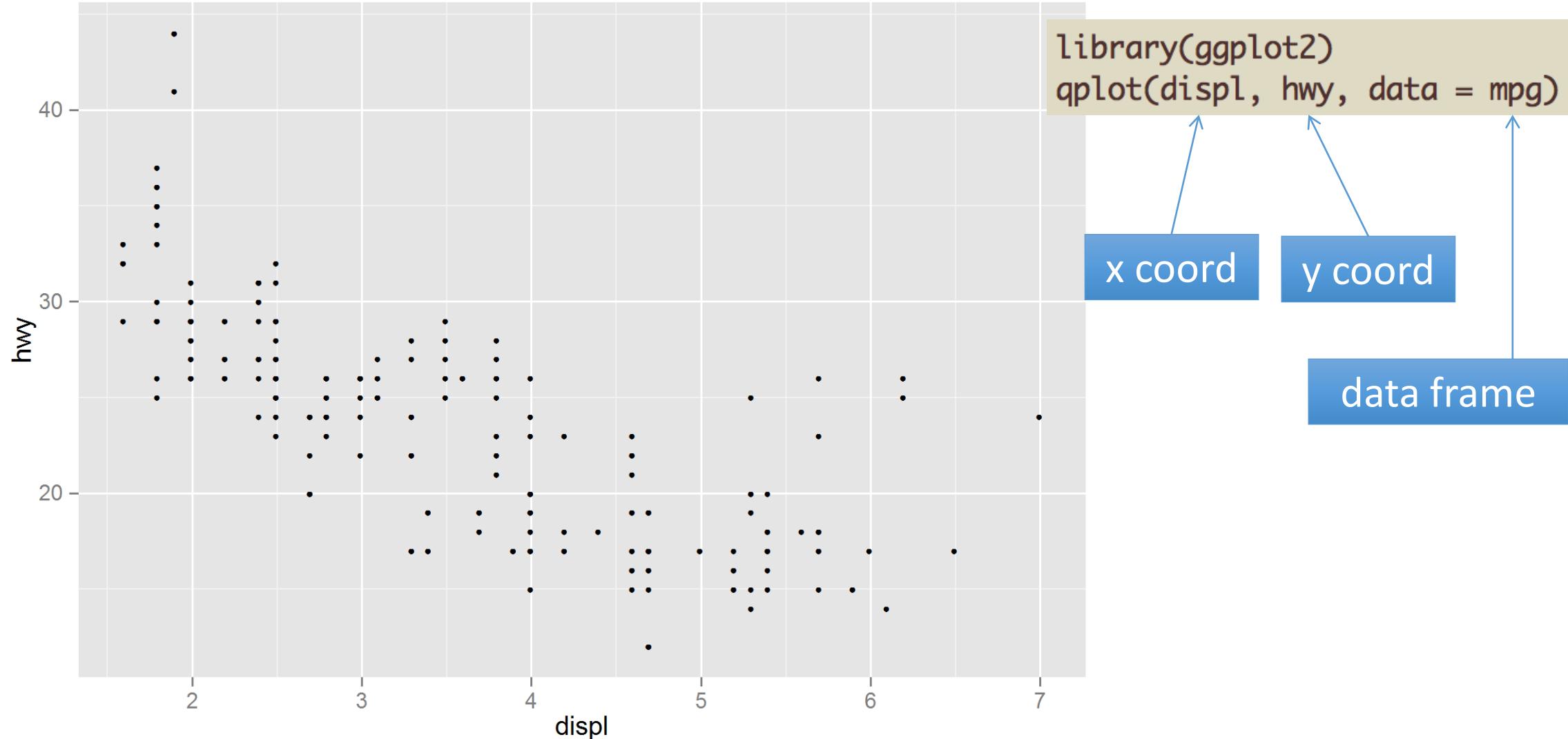
```
> library(ggplot2)
> str(mpg)
```

'data.frame': 234 obs. of 11 variables:

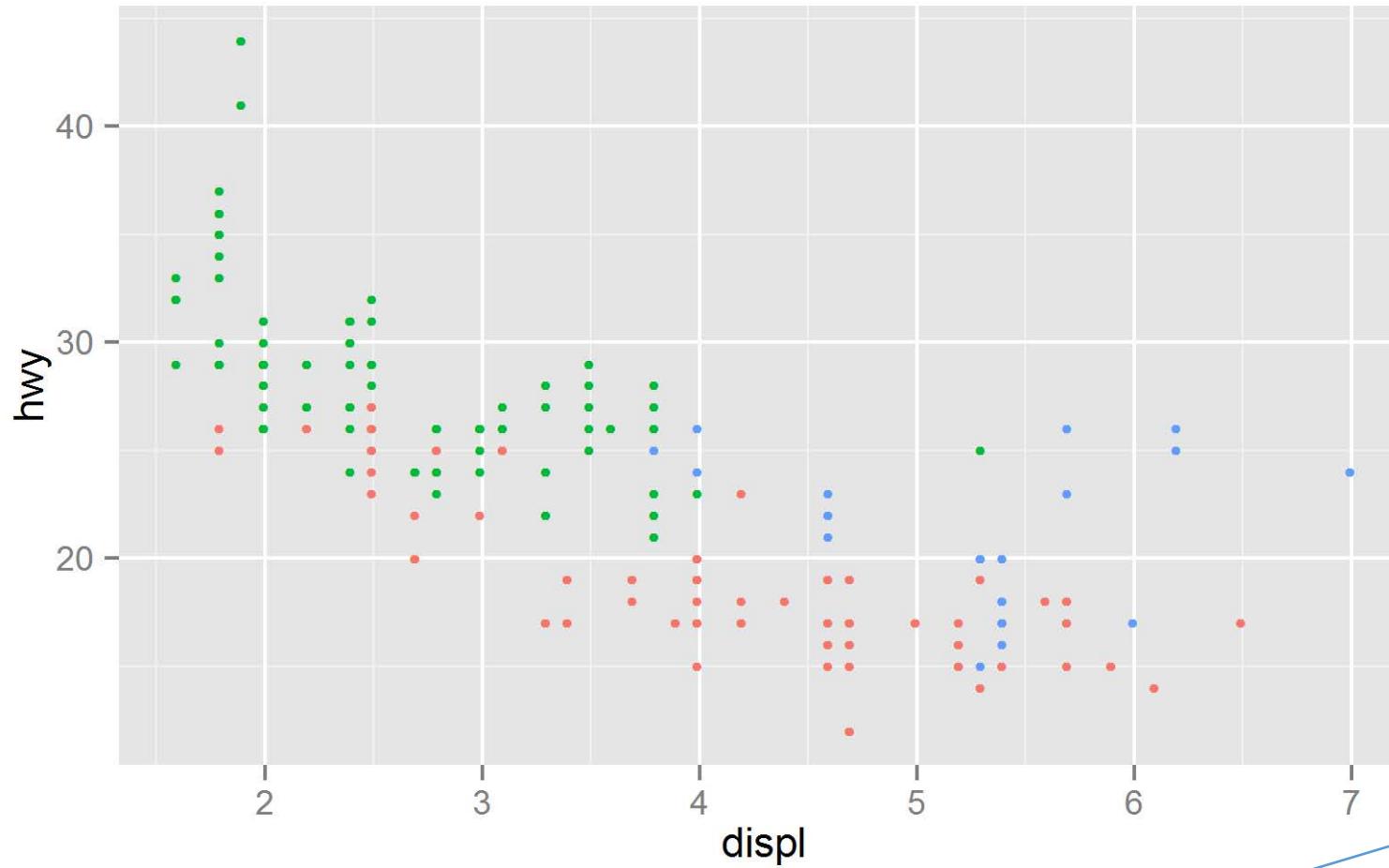
\$ manufacturer	:	Factor w/ 15 levels "audi","chevrolet",...	1 1 1 1 1 1 1 1 1 1 ...
\$ model	:	Factor w/ 38 levels "4runner 4wd",...	2 2 2 2 2 2 2 3 3 3 ...
\$ displ	:	num 1.8 1.8 2 2 2.8 2.8 3.1 1.8 1.8 2 ...	
\$ year	:	int 1999 1999 2008 2008 1999 1999 2008 1999 1999 2008 ...	
\$ cyl	:	int 4 4 4 4 6 6 6 4 4 4 ...	
\$ trans	:	Factor w/ 10 levels "auto(av)","auto(l3)",...	4 9 10 1 4 9 1 9 4 10 ...
...			
\$ drv	:	Factor w/ 3 levels "4","f","r": 2 2 2 2 2 2 2 1 1 1 ...	
\$ cty	:	int 18 21 20 21 16 18 18 18 16 20 ...	
\$ hwy	:	int 29 29 31 30 26 26 27 26 25 28 ...	
\$ fl	:	Factor w/ 5 levels "c","d","e","p",...: 4 4 4 4 4 4 4 4 4 4 ...	
\$ class	:	Factor w/ 7 levels "2seater","compact",...: 2 2 2 2 2 2 2 2 2 2 ...	

Factor label information
important for annotation

ggplot2 “Hello, world!”



Modifying aesthetics



```
qplot(displ, hwy, data = mpg, color = drv)
```

auto legend placement

drv

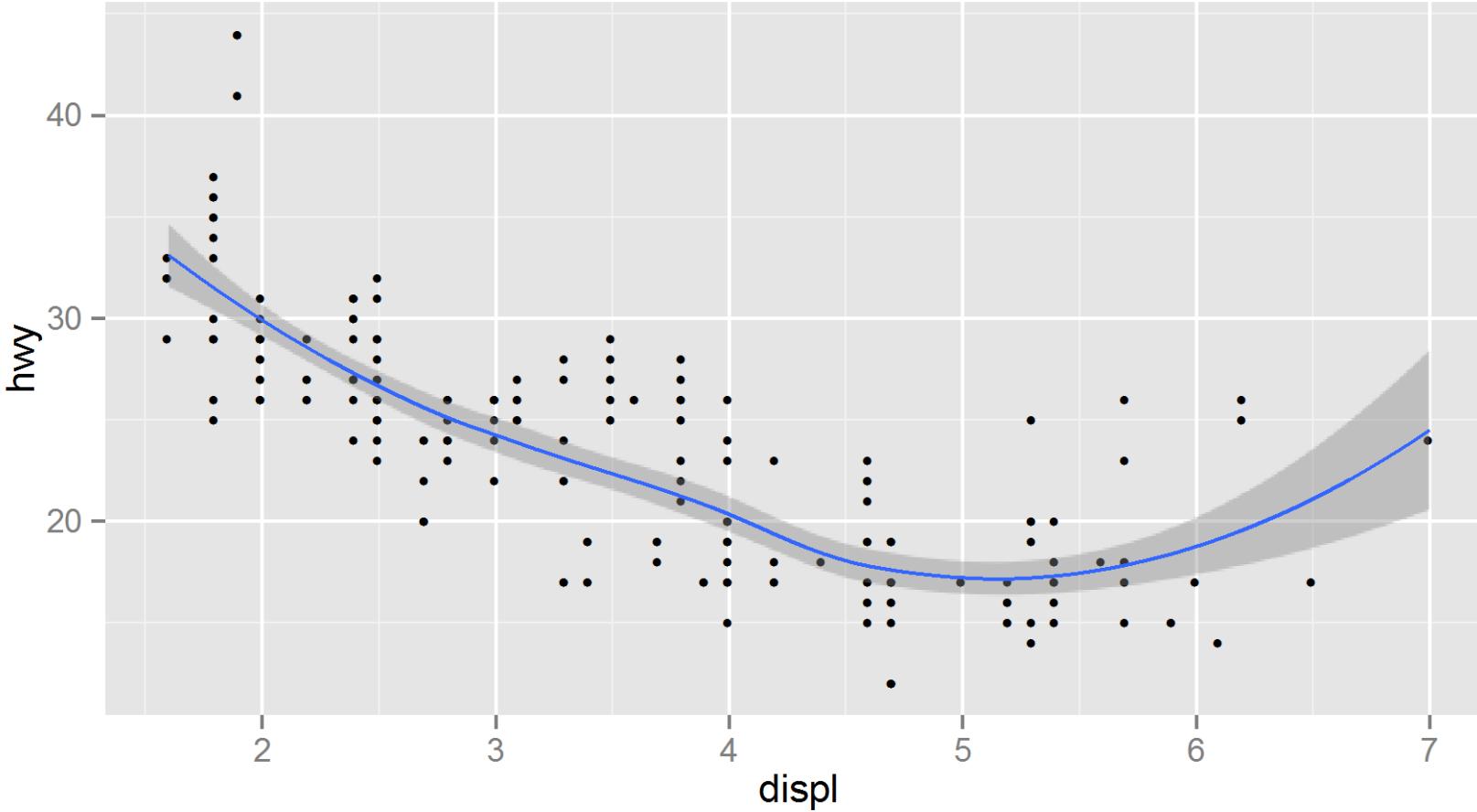
4

f

r

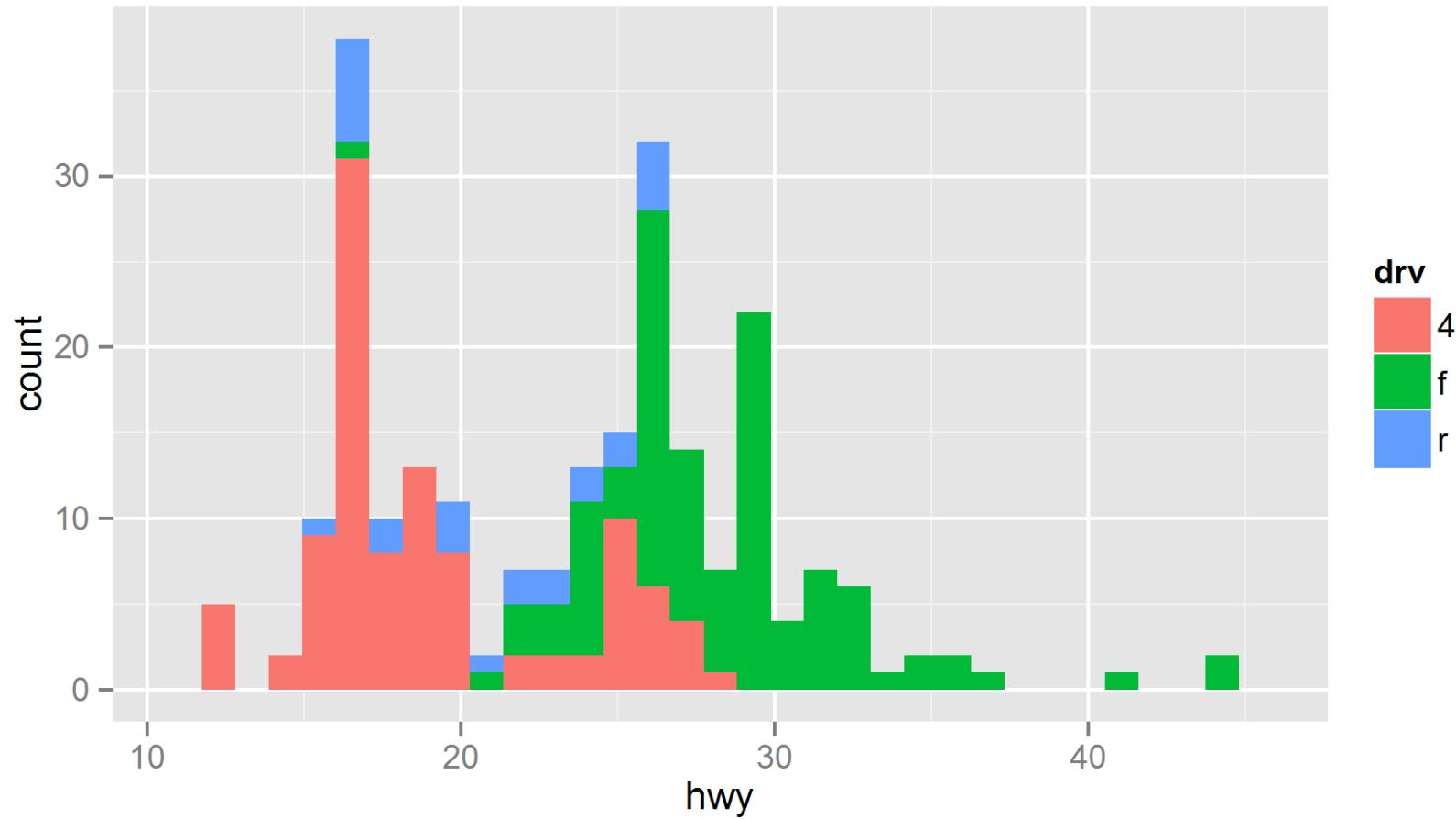
color aesthetic

Adding a geom



```
qplot(displ, hwy, data = mpg, geom = c("point", "smooth"))
```

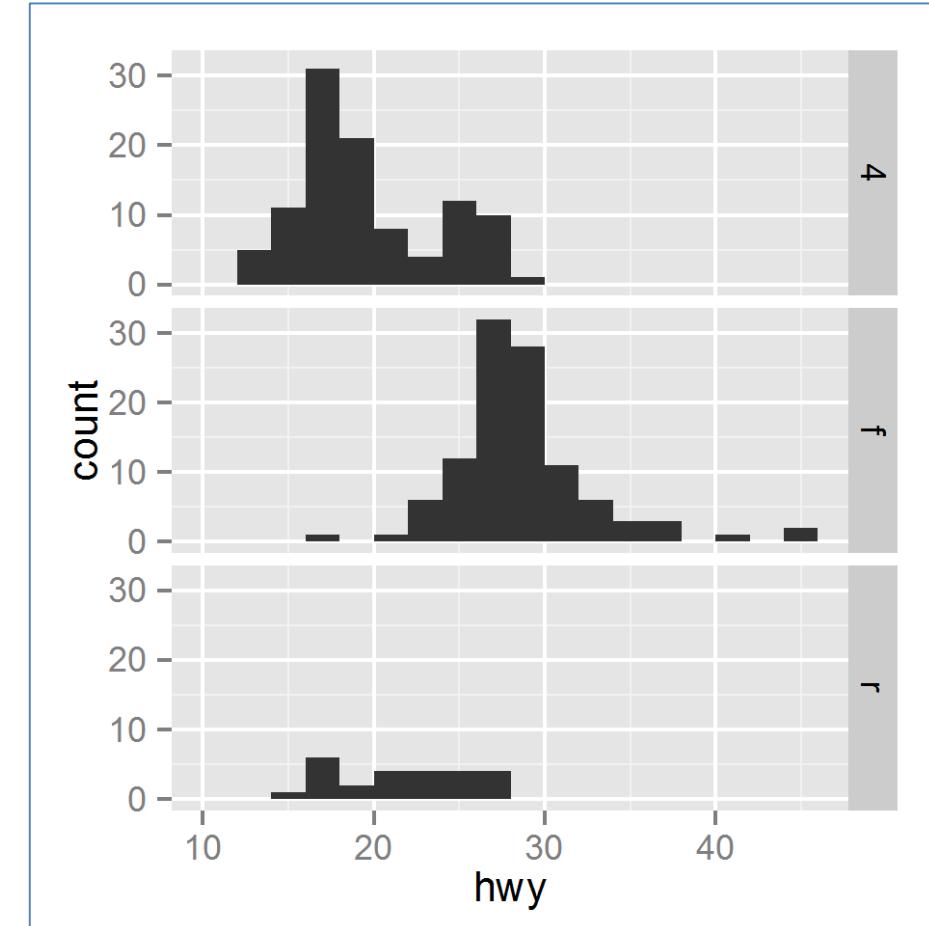
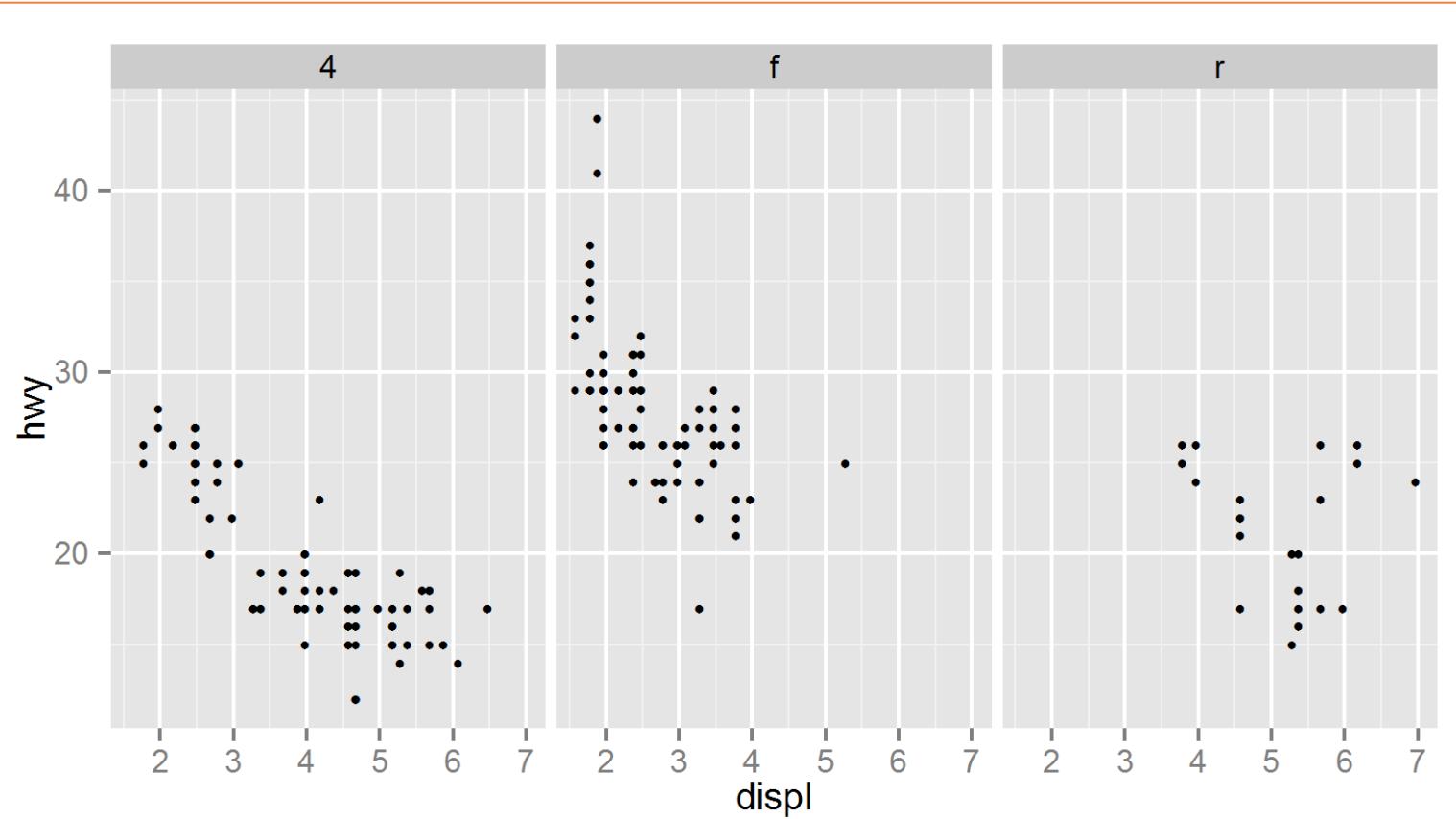
Histograms



qplot(hwy, data = mpg, fill = drv)

Facets

```
qplot(displ, hwy, data = mpg, facets = . ~ drv)
```



```
qplot(hwy, data = mpg, facets = drv ~ ., binwidth = 2)
```

MAACS Cohort

- Mouse Allergen and Asthma Cohort Study
- Baltimore children (aged 5–17)
- Persistent asthma, exacerbation in past year
- Study indoor environment and its relationship with asthma morbidity
- Recent publication: <http://goo.gl/WqE9j8>

Exhaled nitric
oxide

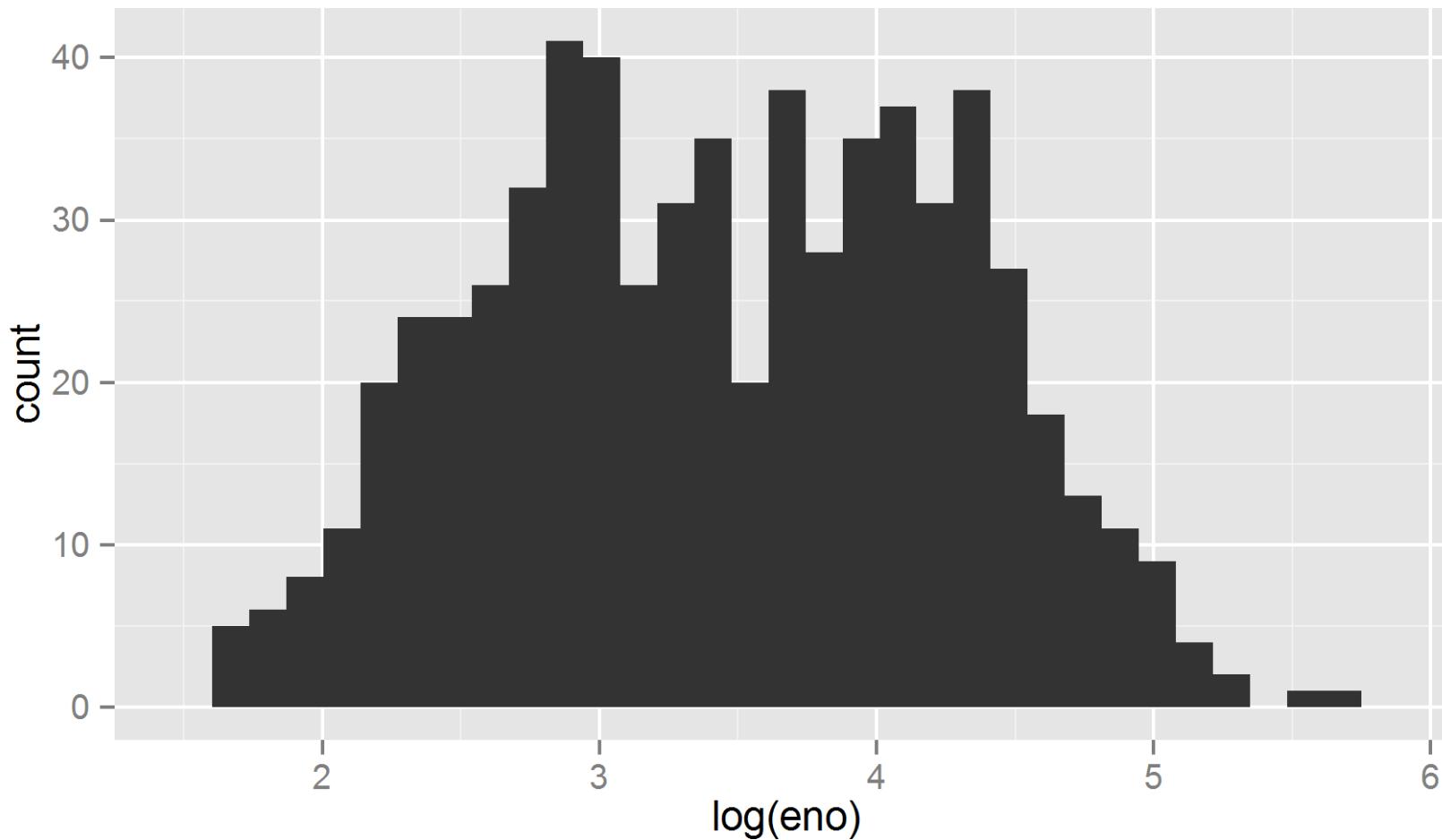
Example: MAACS

```
> str(maacs)
'data.frame': 750 obs. of 5 variables:
 $ id       : int 1 2 3 4 5 6 7 8 9 10 ...
 $ eno      : num 141 124 126 164 99 68 41 50 12 30 ...
 $ duBedMusM: num 2423 2793 3055 775 1634 ...
 $ pm25     : num 15.6 34.4 39 33.2 27.1 ...
 $ _mopos   : Factor w/ 2 levels "no","yes": 2 2 2 2 2 2 2 2 2 2 ...
```

Fine particulate
matter

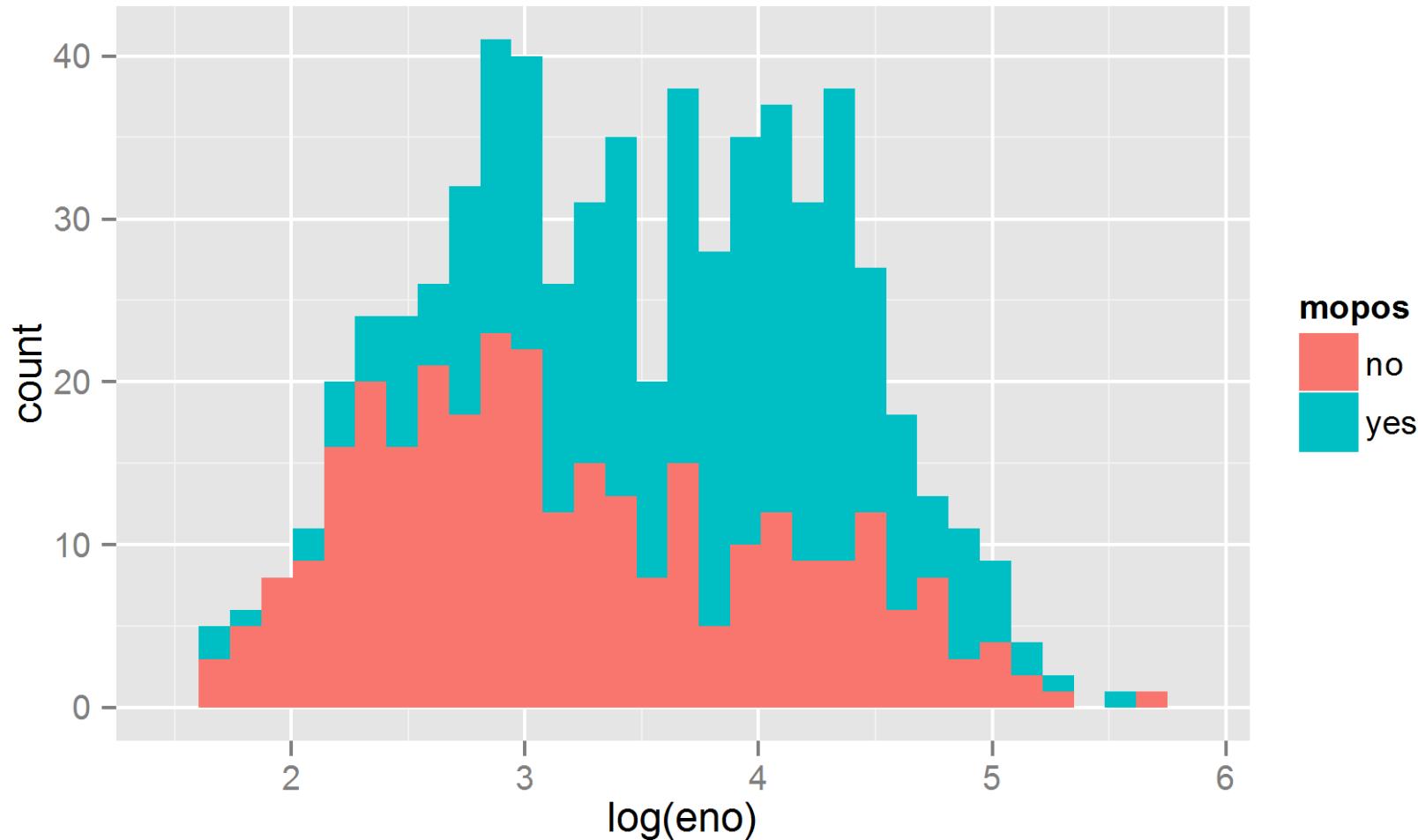
Sensitized to
mouse allergen

Histogram of eNO



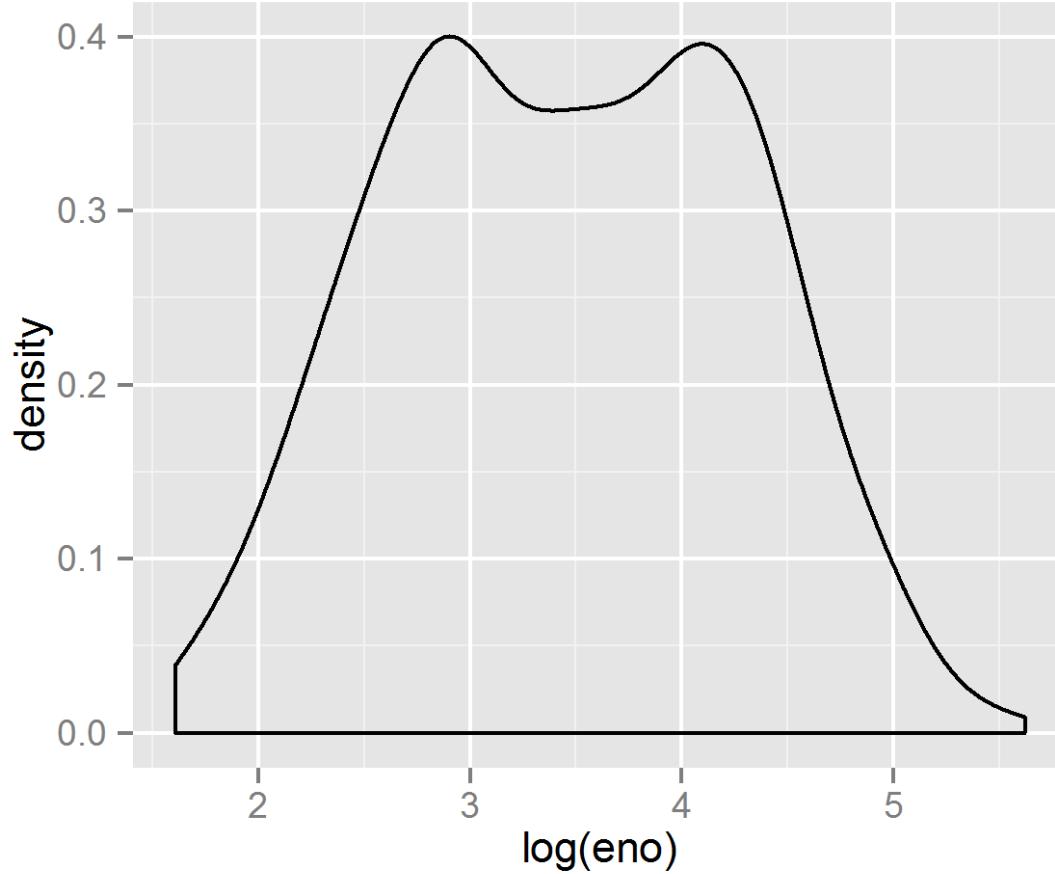
`qplot(log(eno), data = maacs)`

Histogram by Group

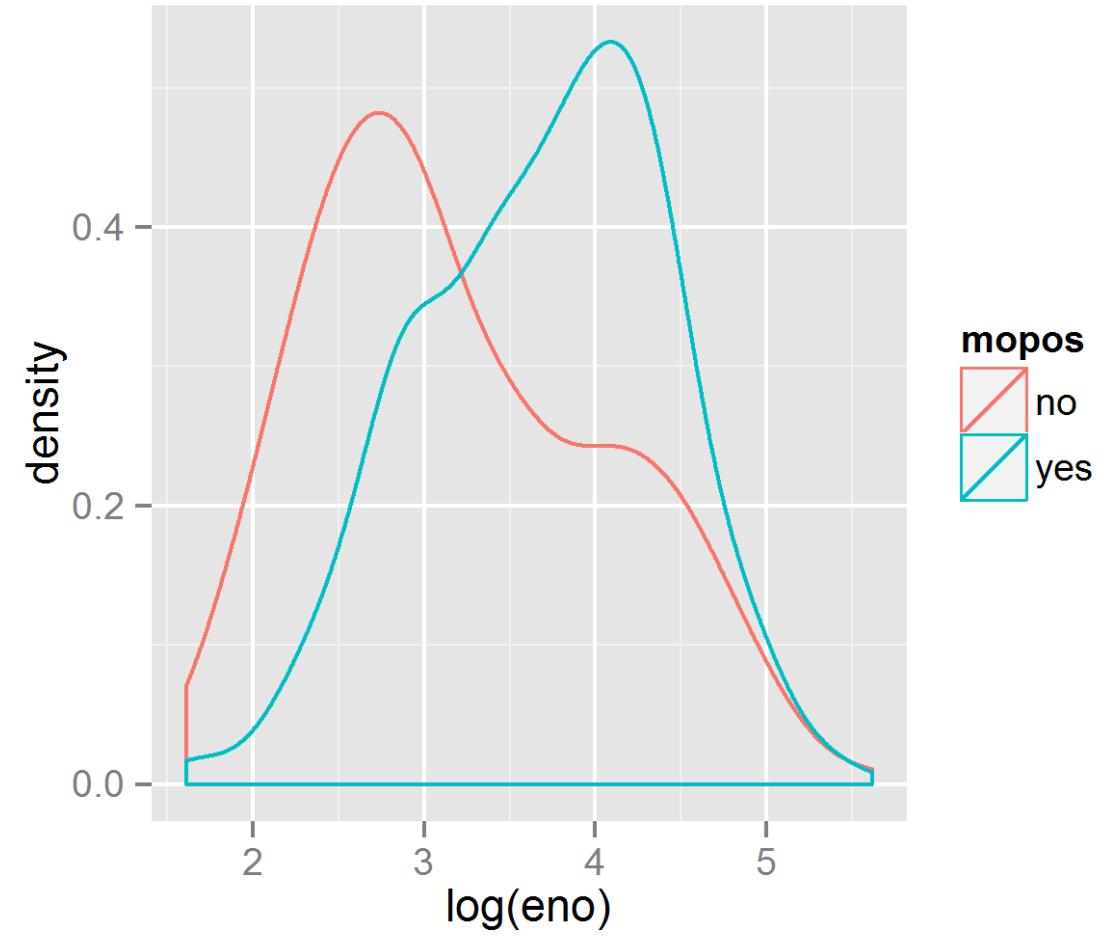


`qplot(log(eno), data = maacs, fill = mopos)`

Density Smooth

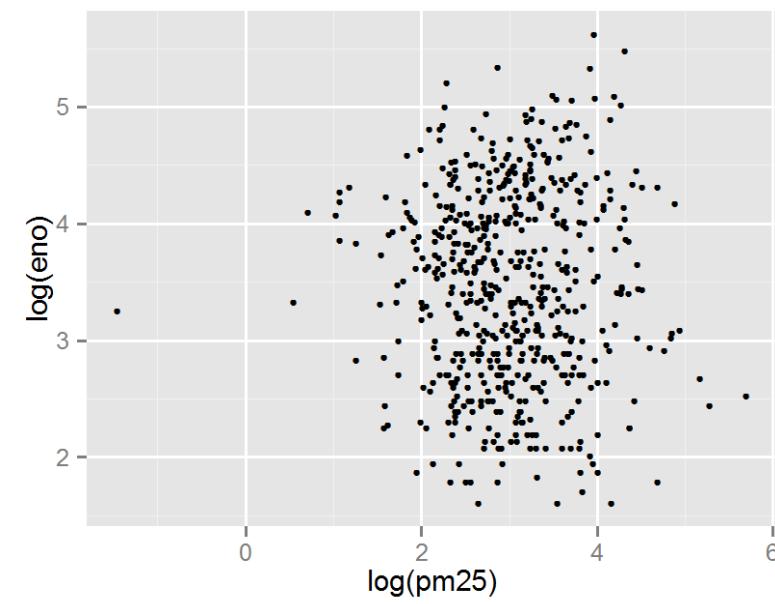


```
qplot(log(eno), data = maacs, geom = "density")
```

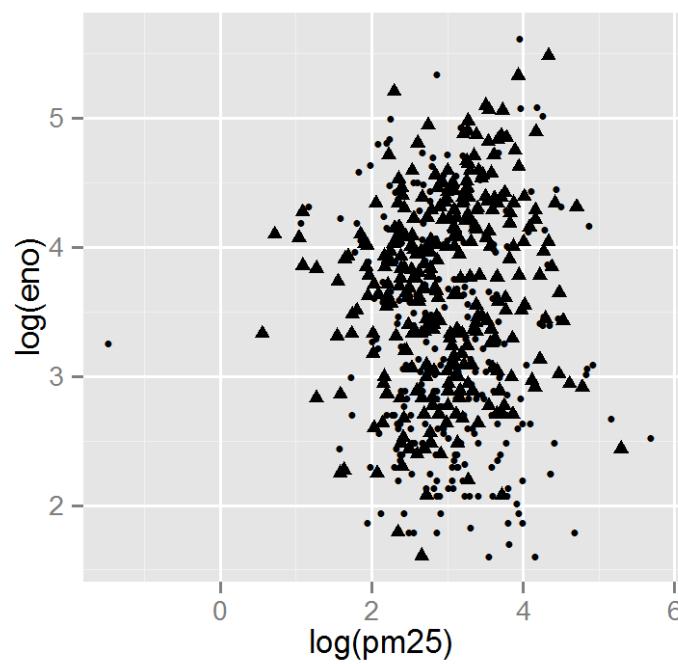


```
qplot(log(eno), data = maacs, geom = "density", color = mpos)
```

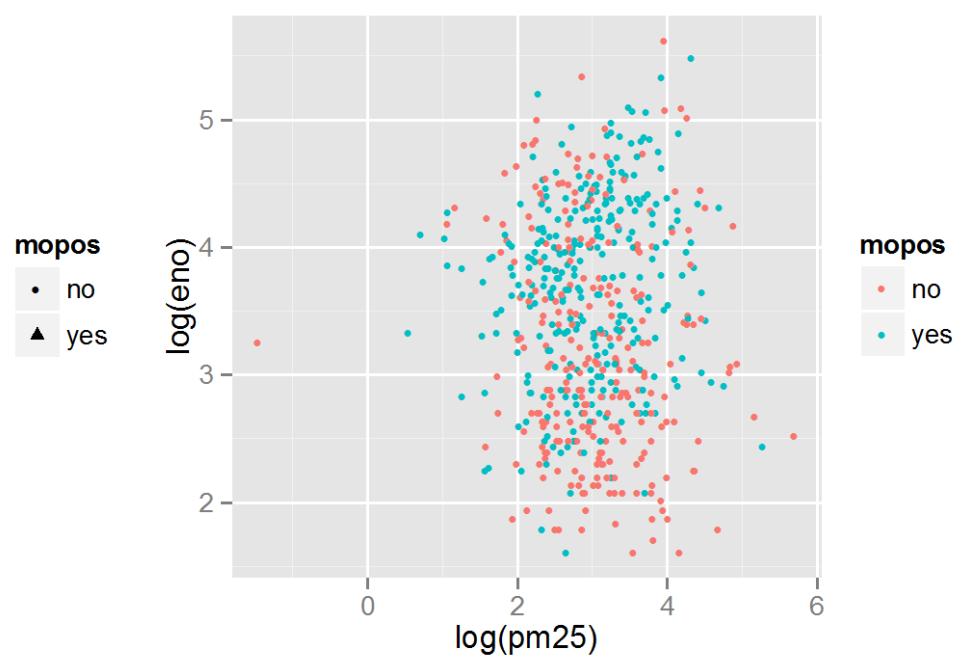
Scatterplots: eNO vs. PM_{2.5}



```
qplot(log(pm25), log(eno), data =  
maacs)
```

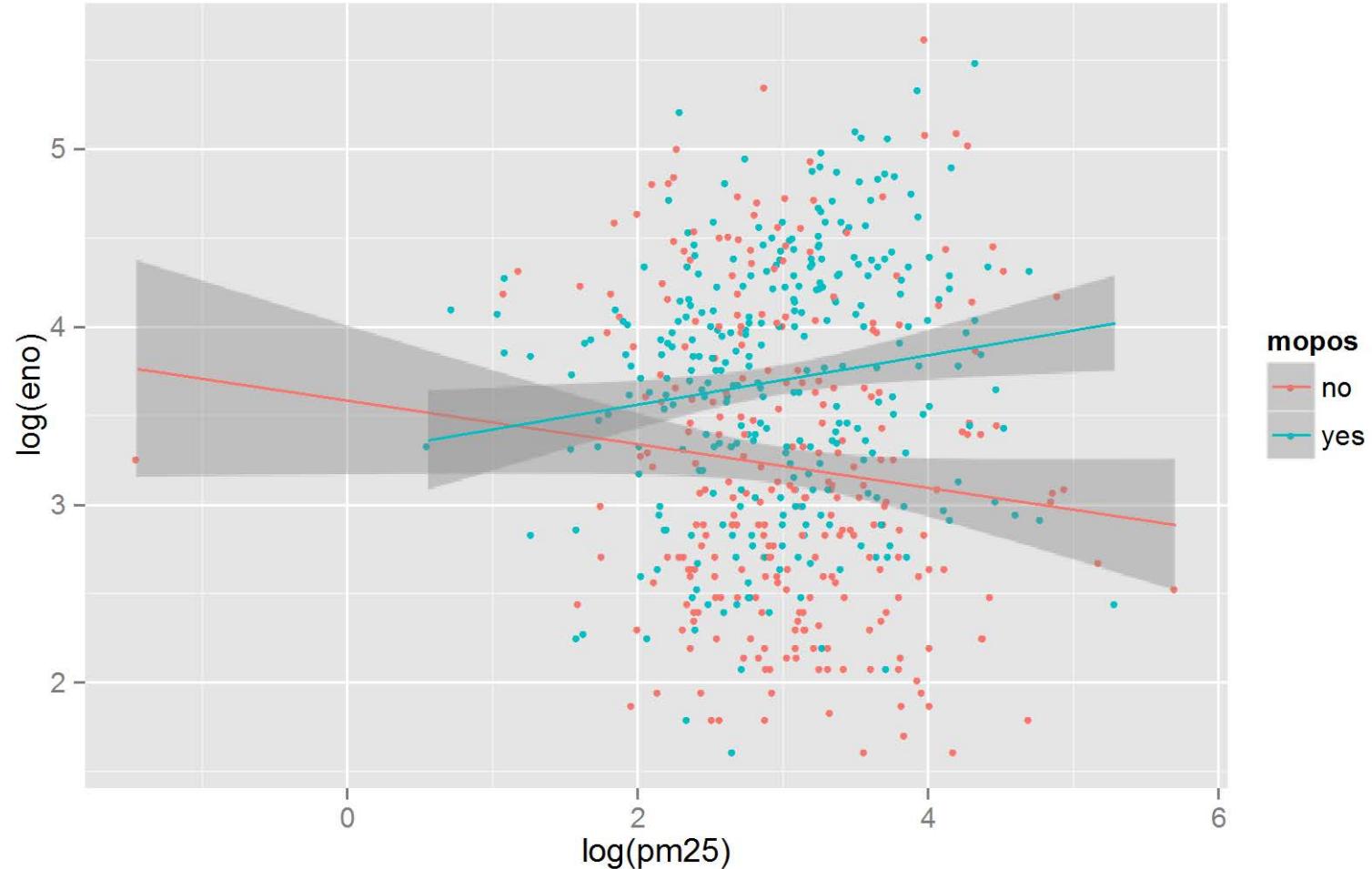


```
qplot(log(pm25), log(eno), data =  
maacs, shape = mpos)
```



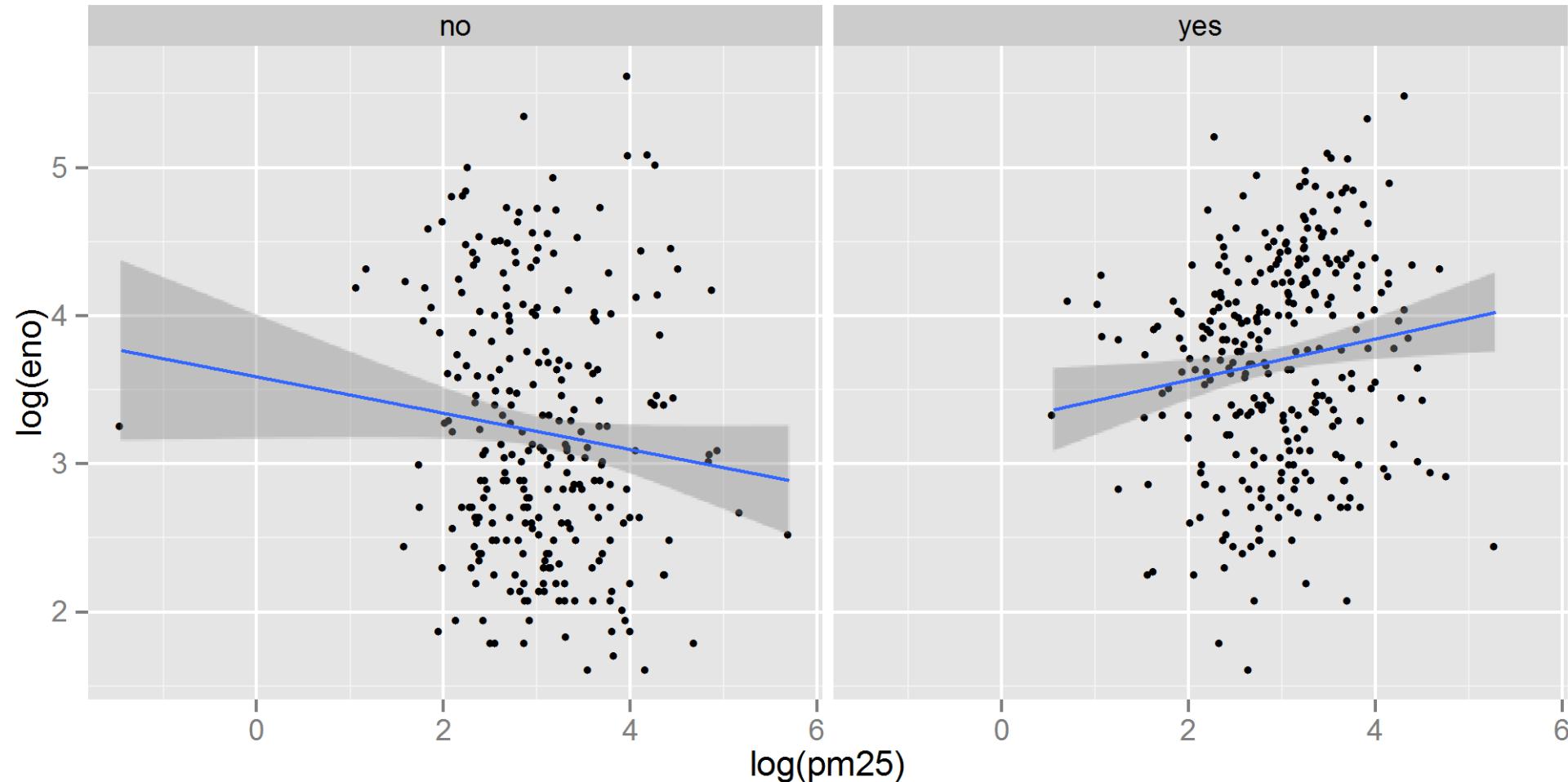
```
qplot(log(pm25), log(eno), data =  
maacs, color = mpos)
```

Scatterplots: eNO vs. PM_{2.5}



```
qplot(log(pm25), log(eno), data = maacs, color = mopos) + geom_smooth(method = "lm")
```

Scatterplots: eNO vs. PM_{2.5}



```
qplot(log(pm25), log(eno), data = maacs, facets = . ~ mopes) + geom_smooth(method = "lm")
```

Summary of qplot()

- The qplot() function is the analog to plot() but with many built-in features
- Syntax somewhere in between base/lattice
- Produces very nice graphics, essentially publication ready (if you like the design)
- Difficult to go against the grain/customize (don't bother; use full ggplot2 power in that case)

Resources

- The *ggplot2* book by Hadley Wickham
- The *R Graphics Cookbook* by Winston Chang (examples in base plots and in ggplot2)
- ggplot2 web site (<http://ggplot2.org>)
- ggplot2 mailing list (<http://goo.gl/OdW3uB>), primarily for developers

Building Plots with ggplot2

- When building plots in ggplot2 (rather than using qplot) the “artist’s palette” model may be the closest analogy
- Plots are built up in layers
 - Plot the data
 - Overlay a summary
 - Metadata and annotation

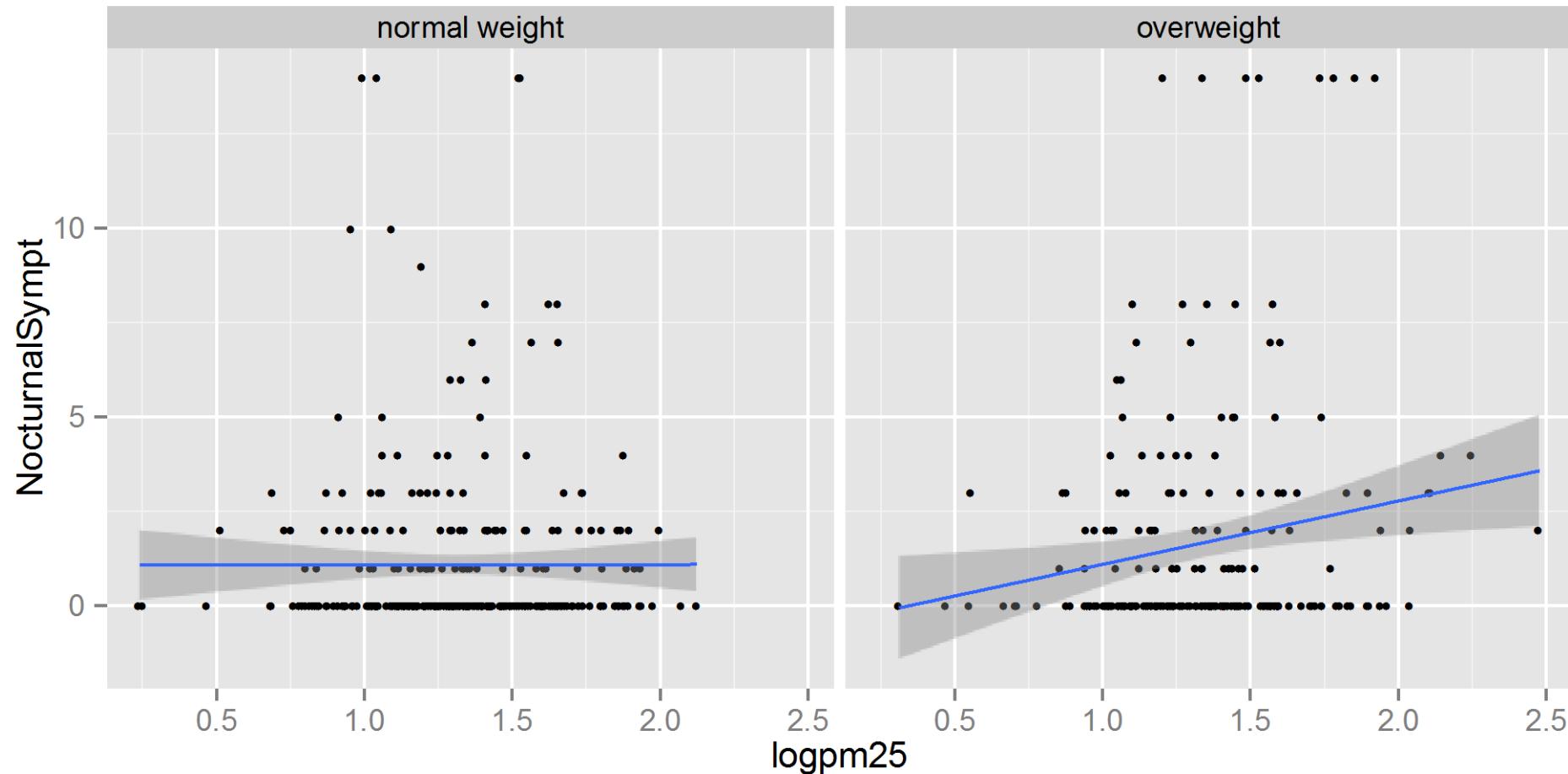
Basic Components of a ggplot2 Plot

- A **data frame**
- **aesthetic mappings**: how data are mapped to color, size
- **geoms**: geometric objects like points, lines, shapes.
- **facets**: for conditional plots.
- **stats**: statistical transformations like binning, quantiles, smoothing.
- **scales**: what scale an aesthetic map uses (example: male = red, female = blue).
- **coordinate system**

Example: BMI, PM_{2.5}, Asthma

- Mouse Allergen and Asthma Cohort Study
- Baltimore children (age 5-17)
- Persistent asthma, exacerbation in past year
- Does BMI (normal vs. overweight) modify the relationship between PM_{2.5} and asthma symptoms?

Basic Plot



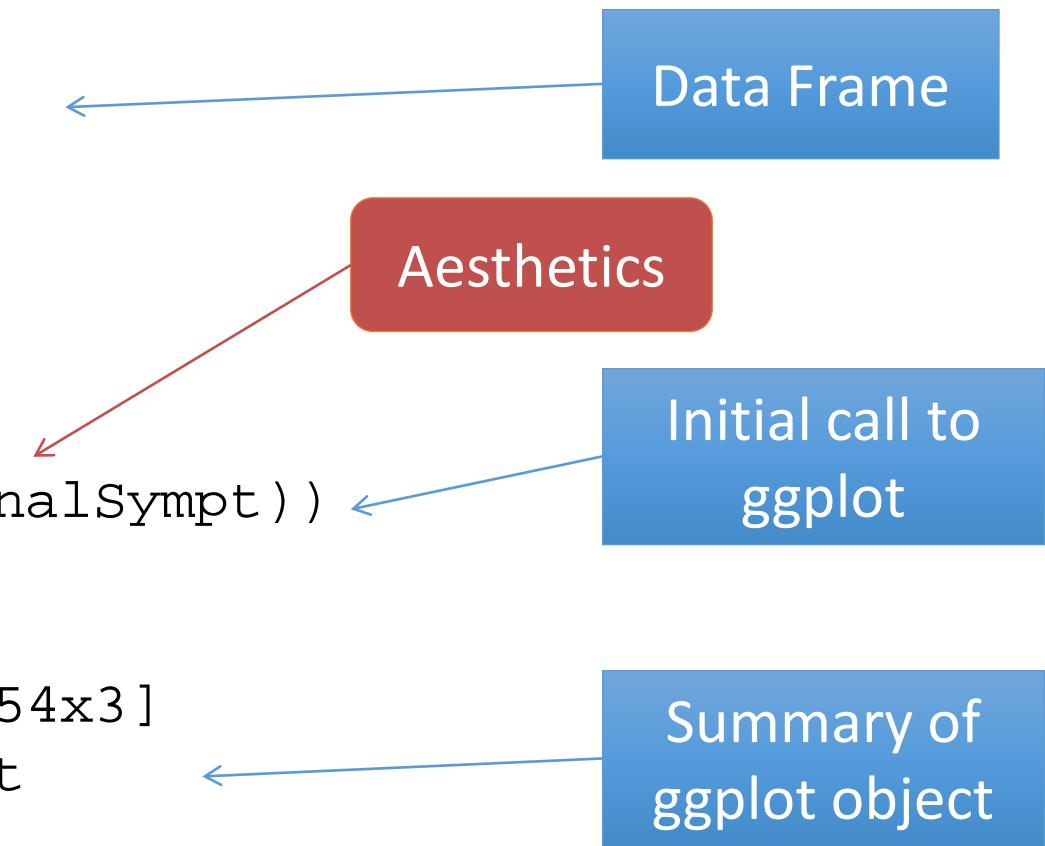
```
qplot(logpm25, NocturnalSympt, data = maacs, facets = . ~ bmicat, geom =  
c("point", "smooth"), method = "lm")
```

Building Up in Layers

```
> head(maacs)
  logpm25      bmicat NocturnalSympt
2 1.5361795 normal    weight        1
3 1.5905409 normal    weight        0
4 1.5217786 normal    weight        0
5 1.4323277 normal    weight        0
6 1.2762320 overweight   overweight     8
8 0.7139103 overweight   overweight     0

> g <- ggplot(maacs, aes(logpm25, NocturnalSympt))

> summary(g)
data: logpm25, bmicat, NocturnalSympt [554x3]
mapping: x = logpm25, y = NocturnalSympt
faceting: facet_null()
```



No Plot Yet!

```
> g <- ggplot(maacs, aes(logpm25, NocturnalSymp) )  
> print(g)  
Error: No layers in plot
```

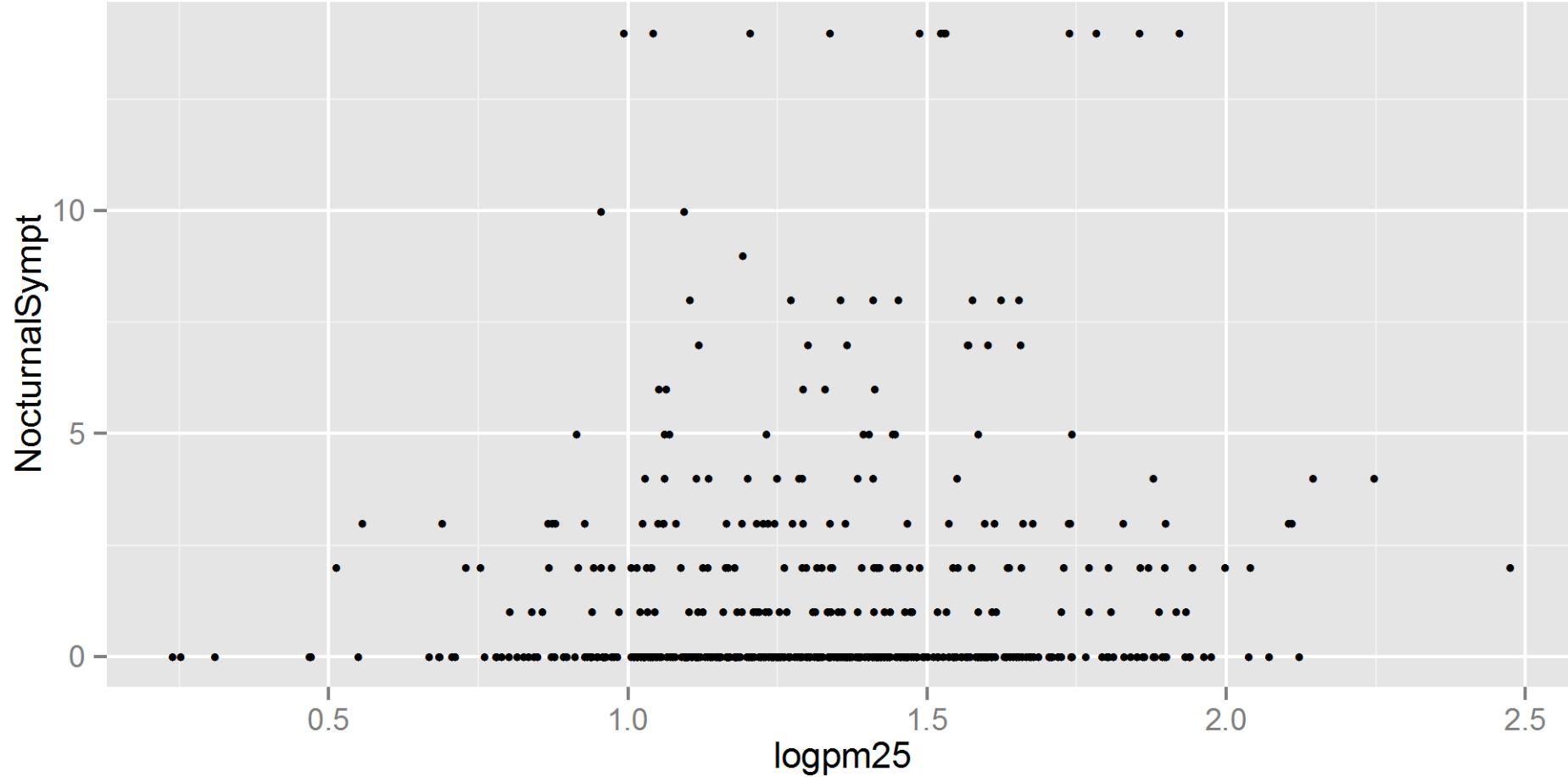
```
> p <- g + geom_point()  
> print(p)
```

Explicitly save and print
ggplot object

```
> g + geom_point()
```

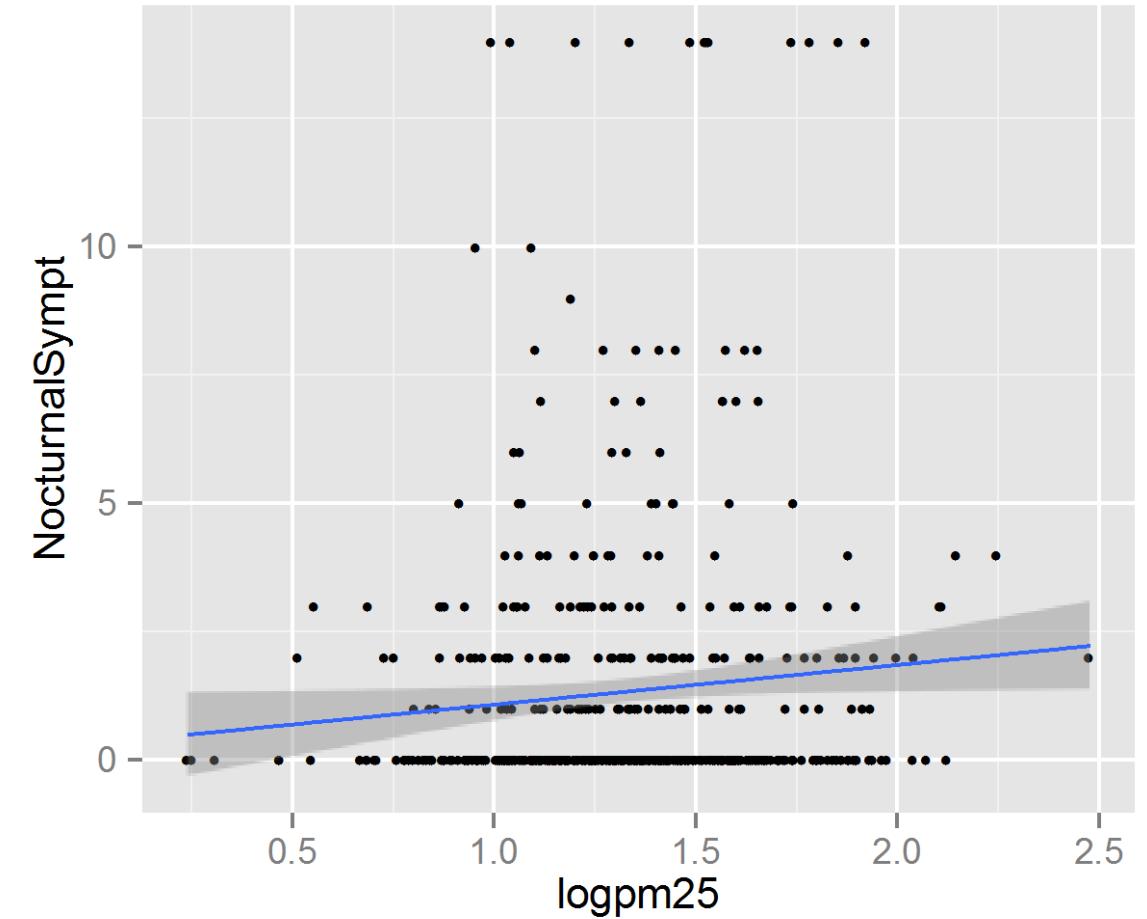
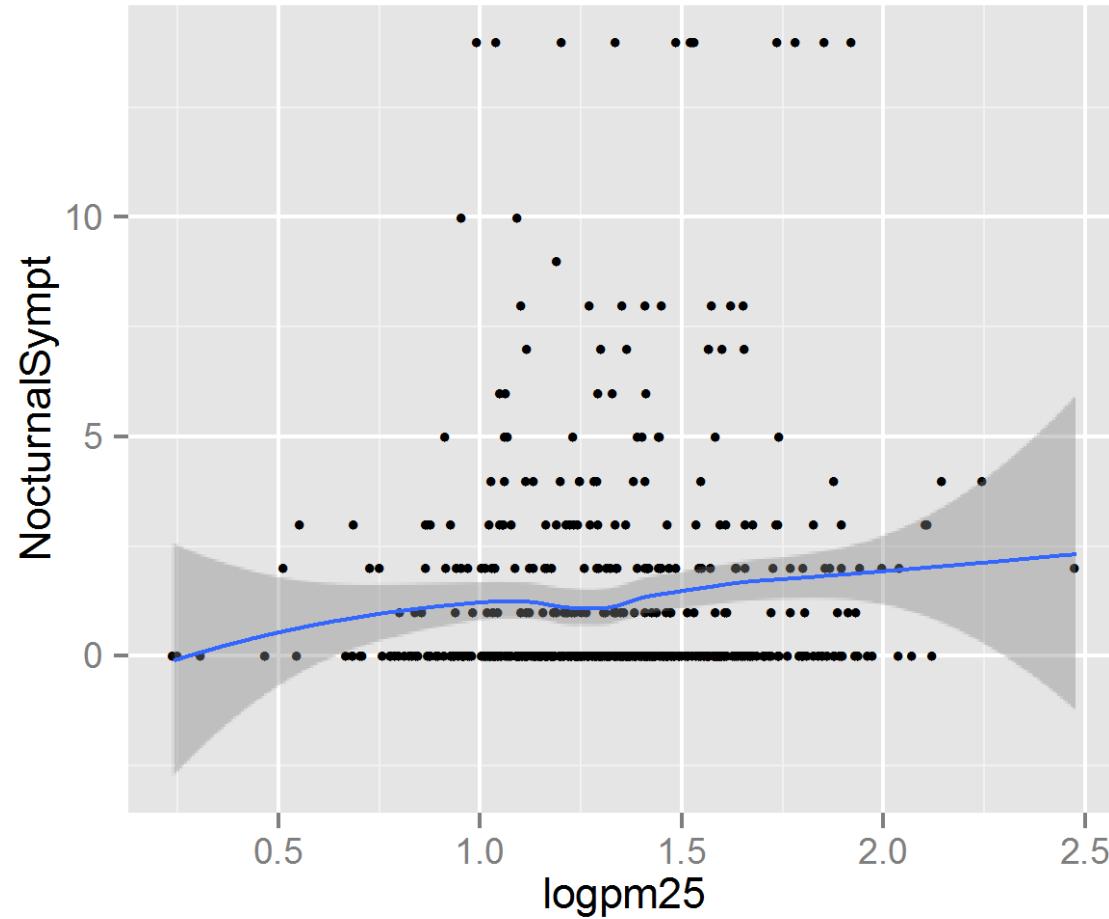
Auto-print plot object
without saving

First Plot with Point Layer



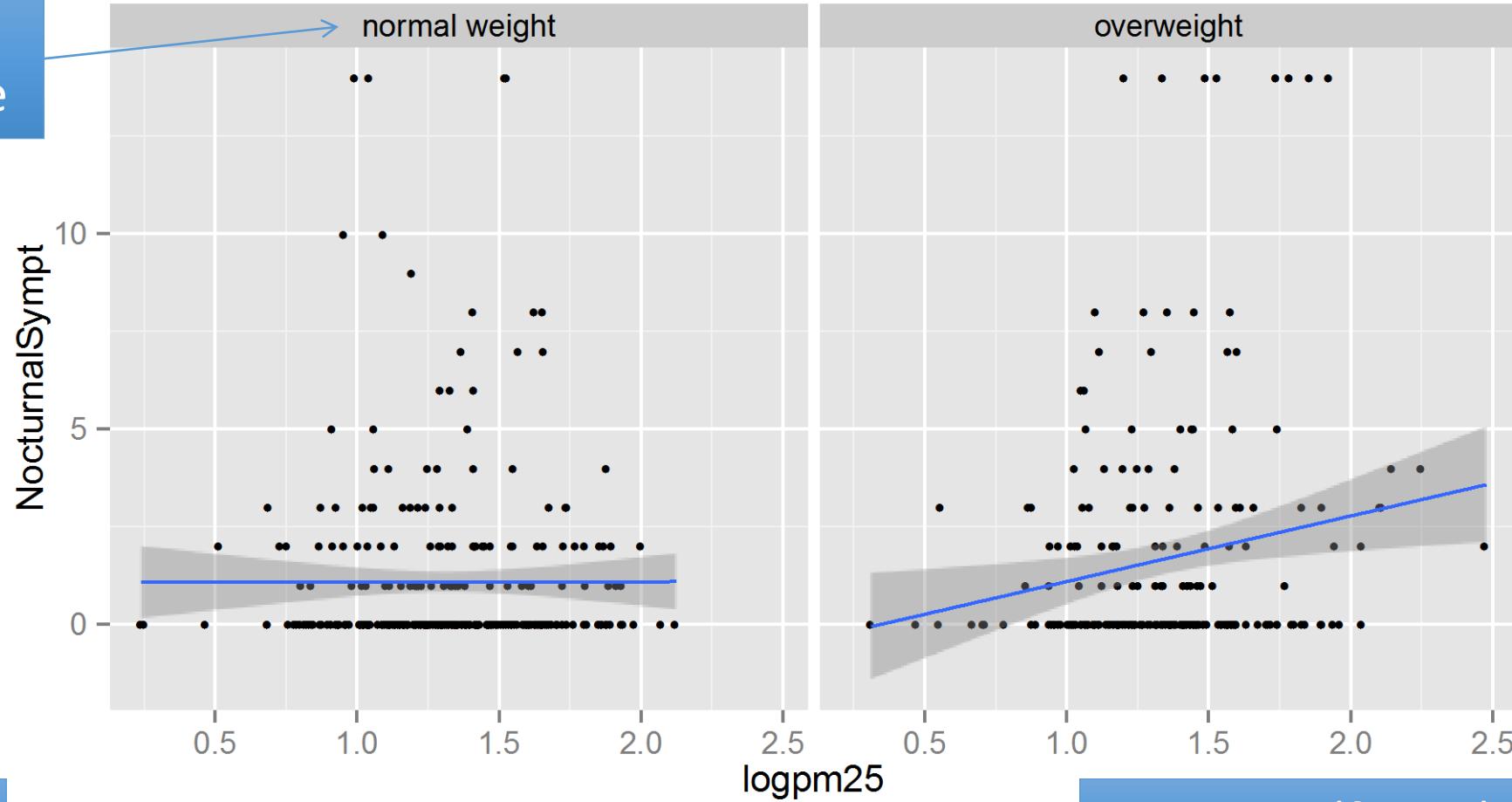
```
g <- ggplot(maacs, aes(logpm25, NocturnalSympt))  
g + geom_point()
```

Adding More Layers: Smooth



Adding More Layers: Facets

Labels from
facet variable



Add facets

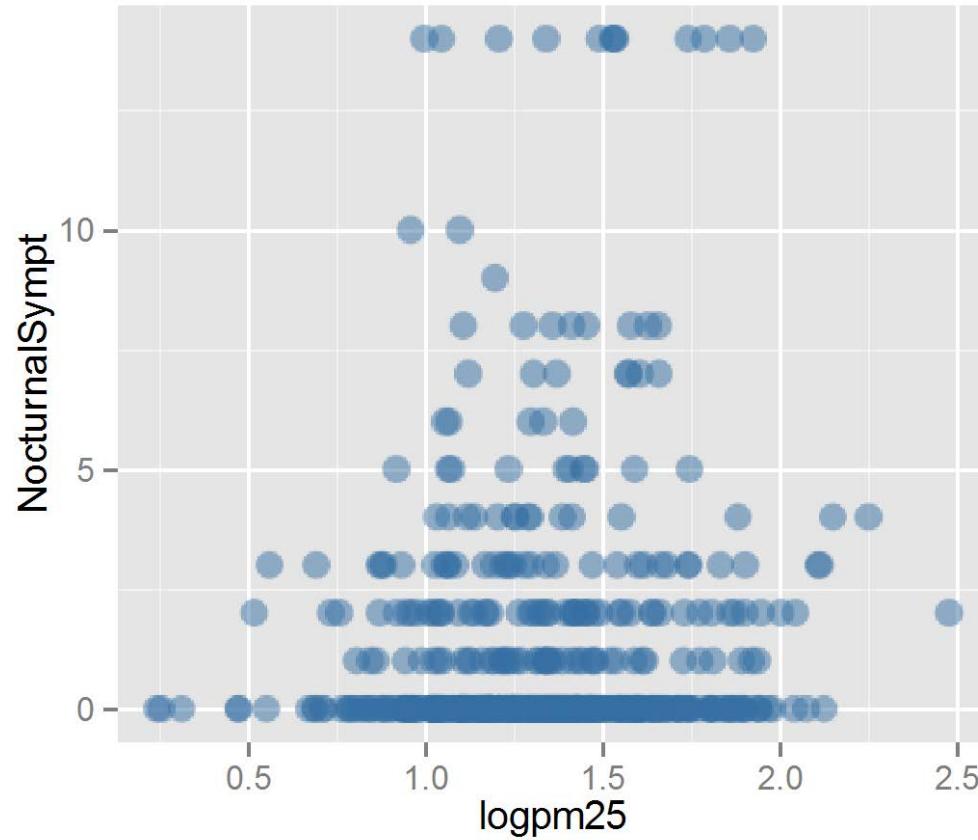
Faceting (factor) variable

```
g + geom_point() + facet_grid(. ~ bmicat) + geom_smooth(method = "lm")
```

Annotation

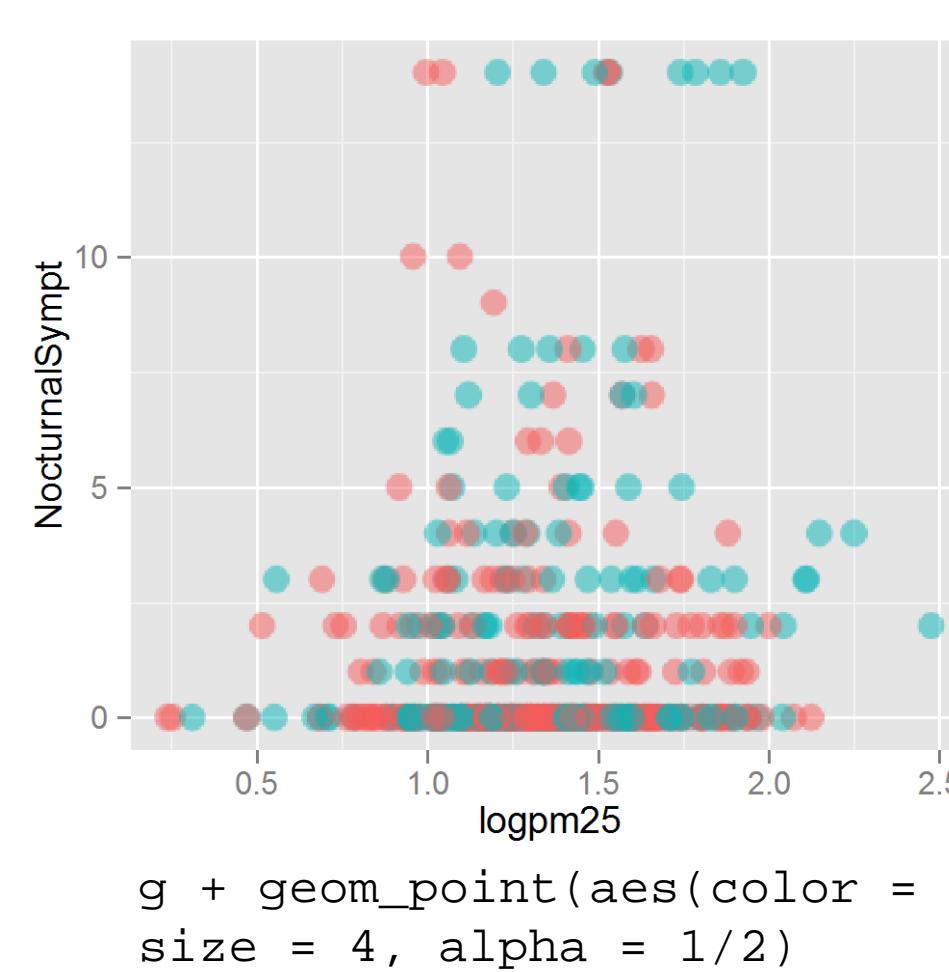
- Labels: `xlab()`, `ylab()`, `labs()`, `ggtitle()`
- Each of the “geom” functions has options to modify
- For things that only make sense globally, use `theme()`
 - Example: `theme(legend.position = "none")`
- Two standard appearance themes are included
 - `theme_gray()`: The default theme (gray background)
 - `theme_bw()`: More stark/plain

Modifying Aesthetics

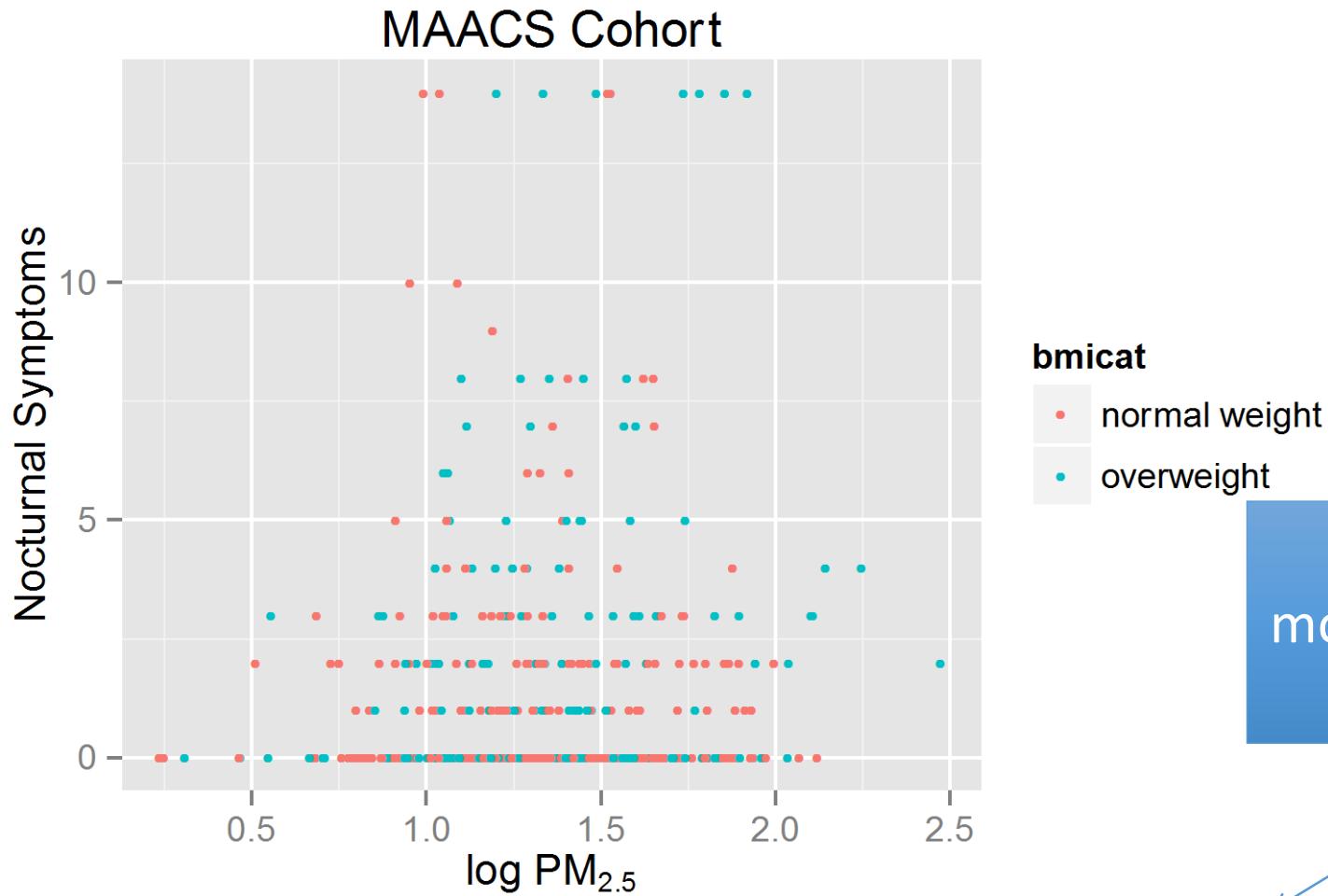


Constant values

```
g + geom_point(color = "steelblue",  
size = 4, alpha = 1/2)
```



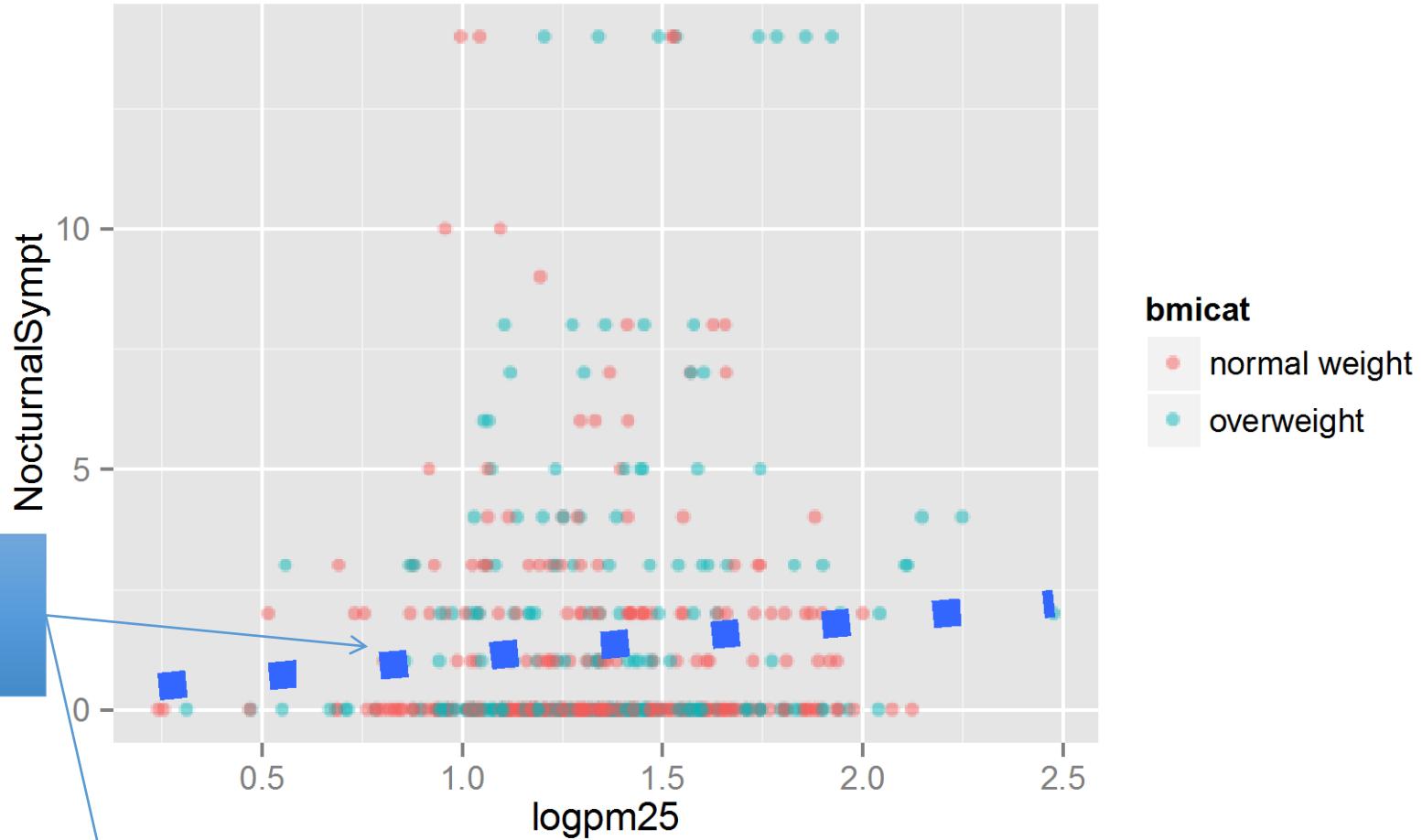
Modifying Labels



labs() function for
modifying titles and x-,
y-axis labels

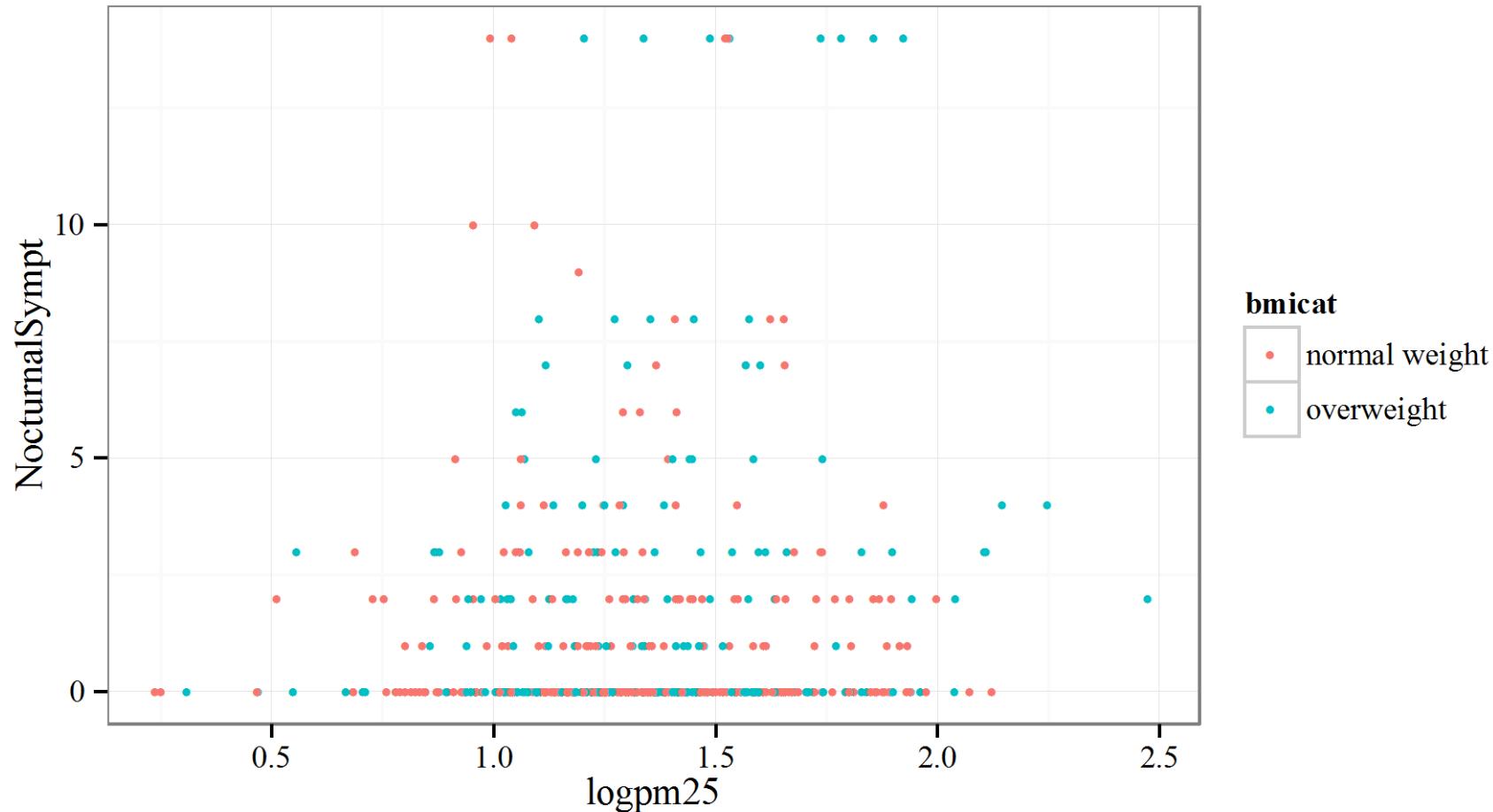
```
g + geom_point(aes(color = bmicat)) + labs(title = "MAACS Cohort") + labs(x = expression("log " * PM[2.5]), y = "Nocturnal Symptoms")
```

Customizing the Smooth



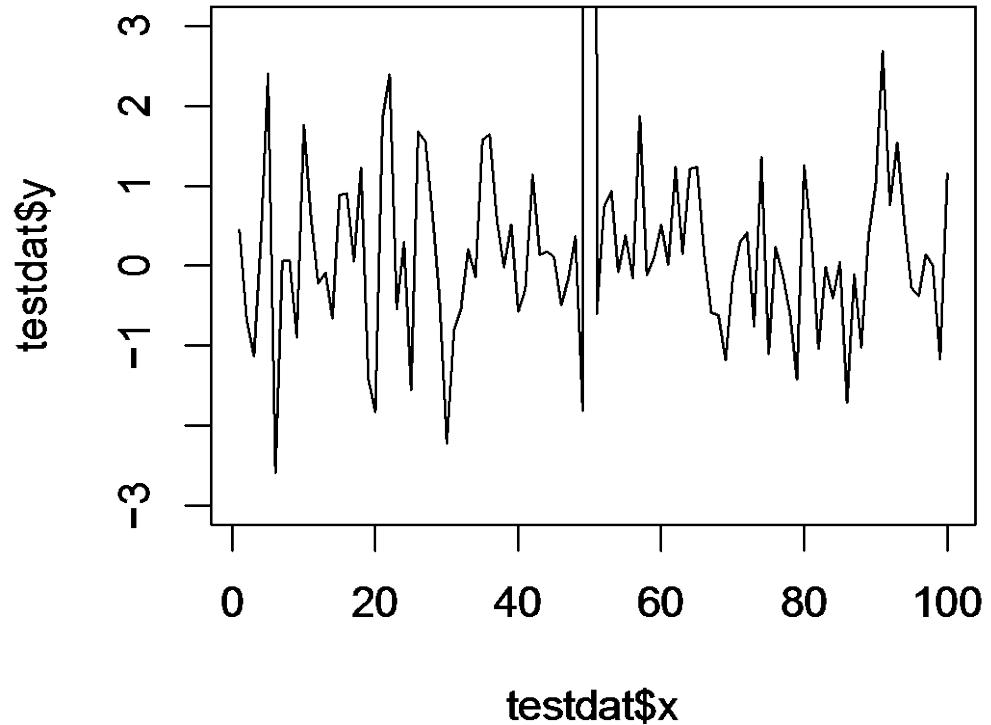
```
g + geom_point(aes(color = bmicat), size = 2, alpha = 1/2) +  
  geom_smooth(size = 4, linetype = 3, method = "lm", se = FALSE)
```

Changing the Theme

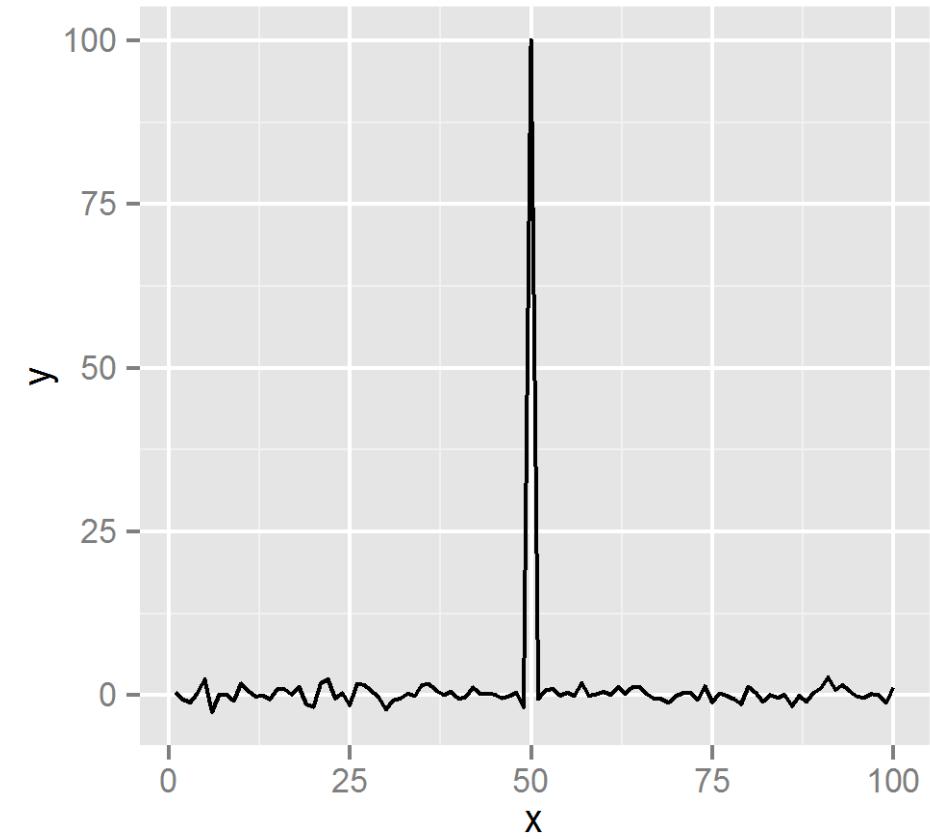


```
g + geom_point(aes(color = bmicat)) + theme_bw(base_family = "Times")
```

A Notes about Axis Limits

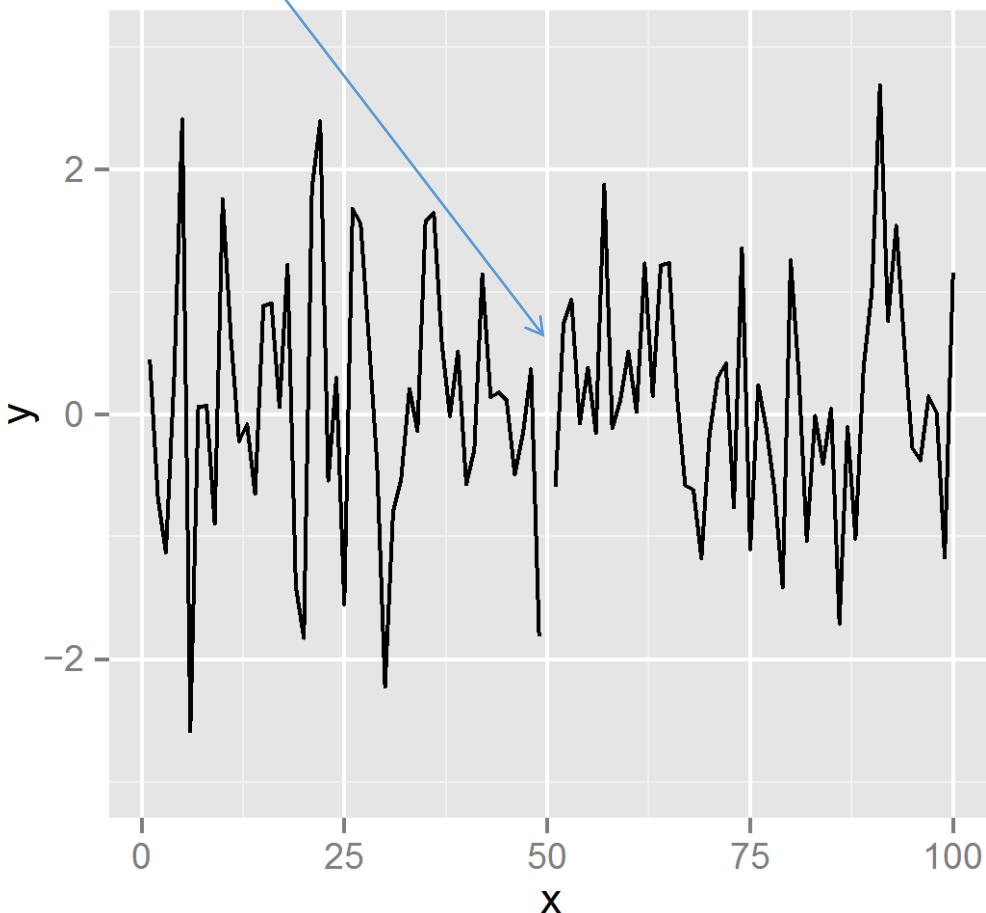


```
testdat <- data.frame(x = 1:100, y = rnorm(100))
testdat[50,2] <- 100 ## Outlier!
plot(testdat$x, testdat$y, type = "l", ylim = c(-3,3))
```



```
g <- ggplot(testdat, aes(x = x, y = y))
g + geom_line()
```

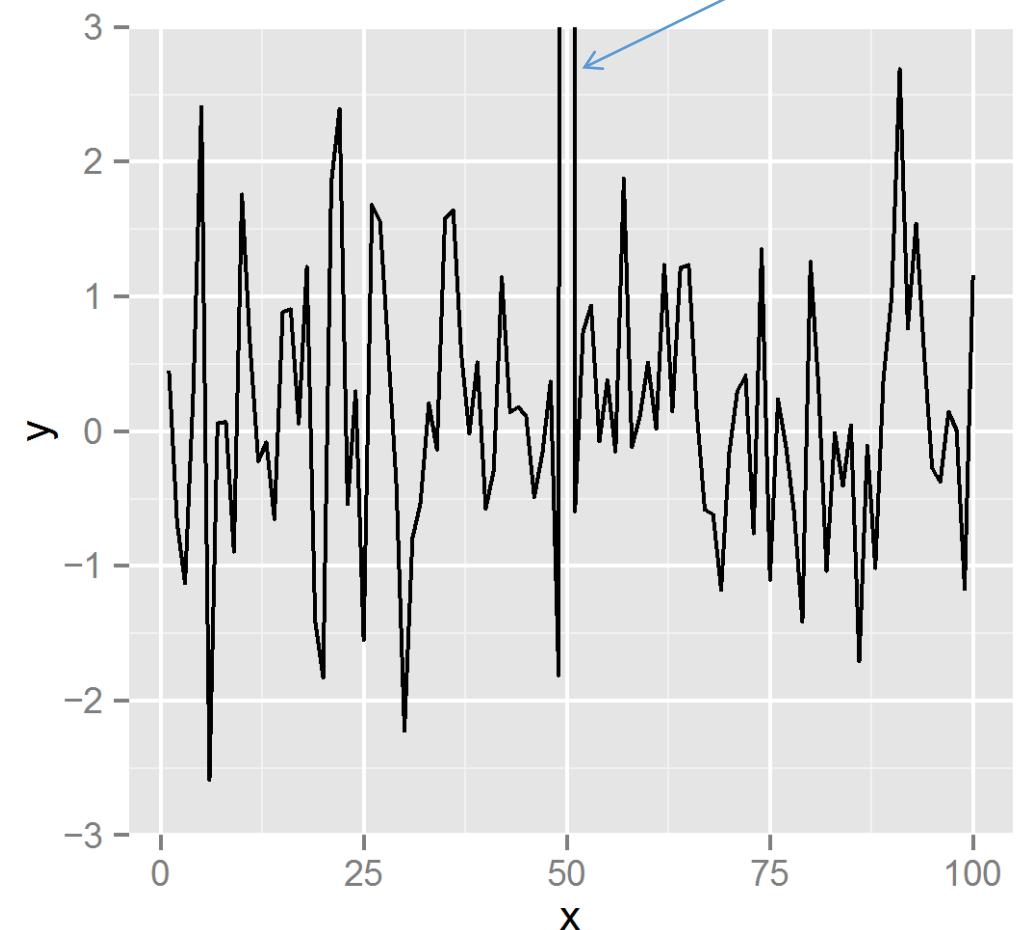
Outlier missing



```
g + geom_line() + ylim(-3, 3)
```

Axis Limits

Outlier included



```
g + geom_line() + coord_cartesian(ylim = c(-3, 3))
```

More Complex Example

- How does the relationship between $\text{PM}_{2.5}$ and nocturnal symptoms vary by BMI and NO_2 ?
- Unlike our previous BMI variable, NO_2 is continuous
- We need to make NO_2 categorical so we can condition on it in the plotting
 - Use the `cut()` function for this

Making NO₂ Deciles

```
## Calculate the deciles of the data
> cutpoints <- quantile(maacs$logno2_new, seq(0, 1, length = 11), na.rm = TRUE)

## Cut the data at the deciles and create a new factor variable
> maacs$no2dec <- cut(maacs$logno2_new, cutpoints)

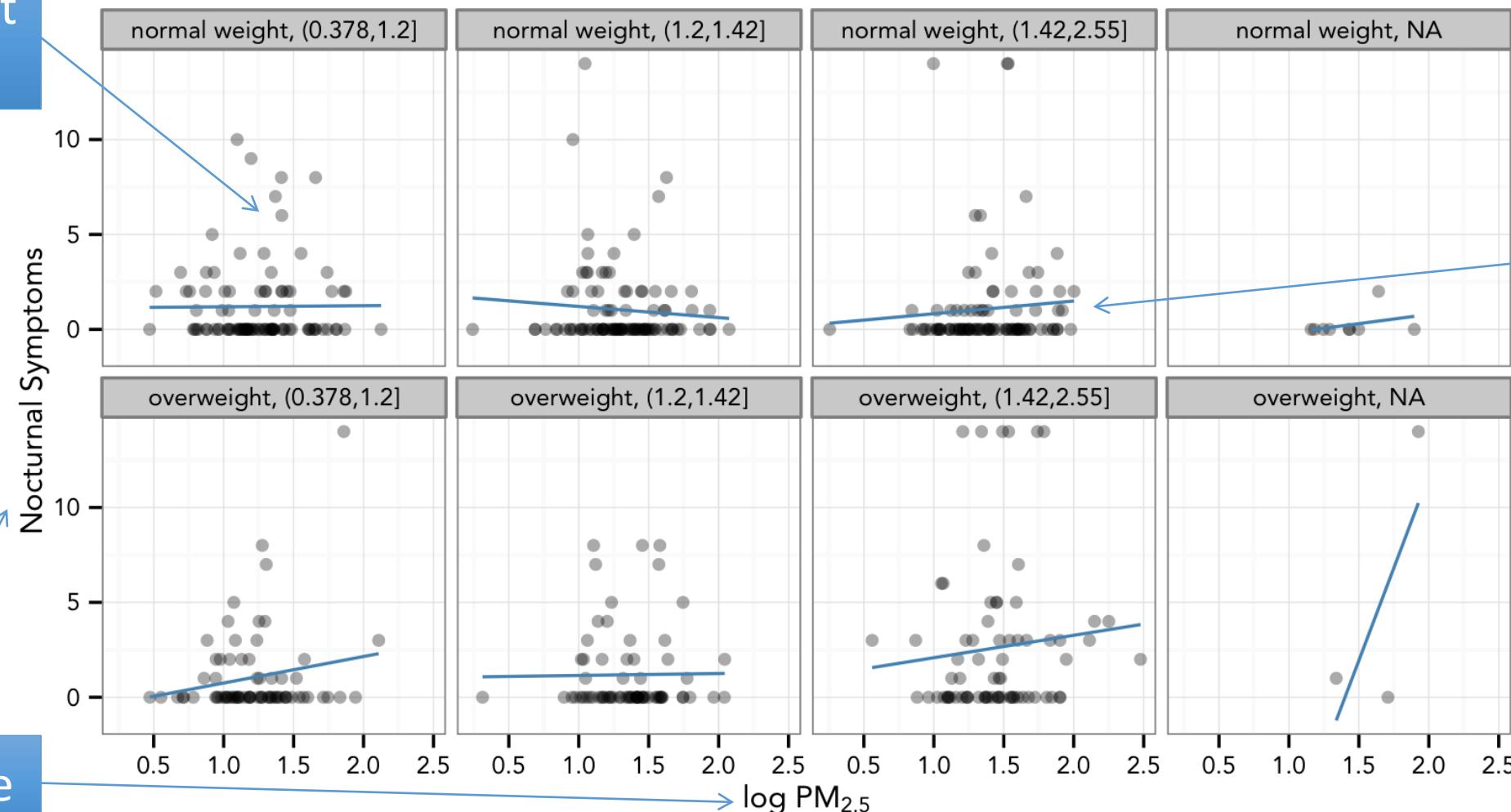
## See the levels of the newly created factor variable
> levels(maacs$no2dec)
[1] "(0.378,0.969]"  "(0.969,1.1]"   "(1.1,1.17]"    "(1.17,1.26]"
[5] "(1.26,1.32]"   "(1.32,1.38]"  "(1.38,1.44]"   "(1.44,1.54]"
[9] "(1.54,1.69]"   "(1.69,2.55]"
```

Non-default font

Final Plot

Multiple panels

Transparent points



Labels/Title

Code for Final Plot

```
## Setup ggplot with data frame
g <- ggplot(maacs, aes(logpm25, NocturnalSymp) )  
  
## Add layers
g + geom_point(alpha = 1/3)
+ facet_wrap(bmicat ~ no2dec, nrow = 2, ncol = 4)
+ geom_smooth(method="lm", se=FALSE, col="steelblue" )
+ theme_bw(base_family = "Avenir", base_size = 10)
+ labs(x = expression("log " * PM[ 2 . 5 ] ) )
+ labs(y = "Nocturnal Symptoms" )
+ labs(title = "MAACS Cohort" )
```

Add points

Make panels

Add smoother

Change theme

Add labels

Summary

- ggplot2 is very powerful and flexible if you learn the “grammar” and the various elements that can be tuned/modified
- Many more types of plots can be made; explore and mess around with the package (references mentioned in Part 1 are useful)

Resources for plotting in R

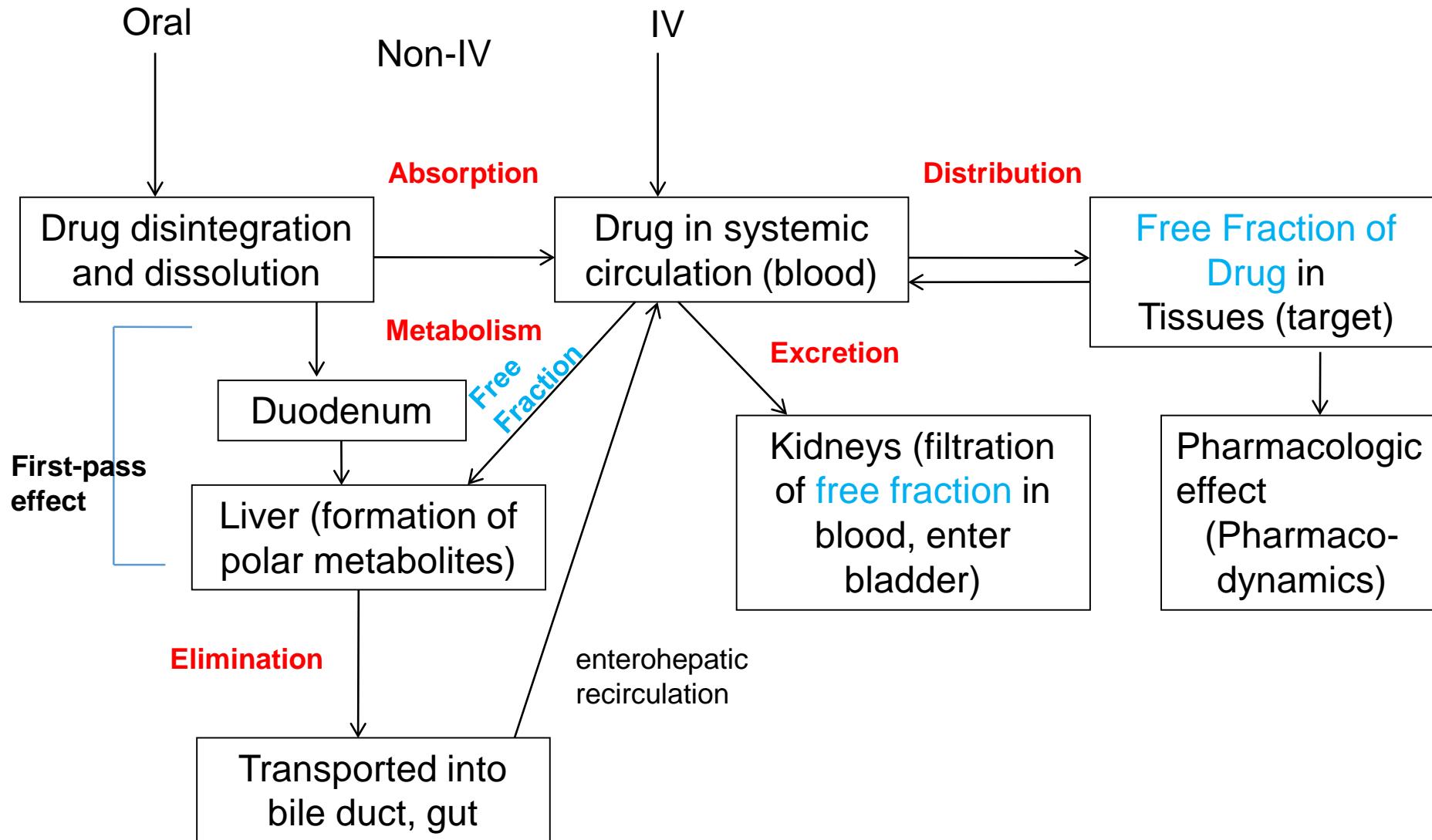
- <http://ggplot2.org/book/>
- <http://cookbook-r.com/Graphs/>

AM Break

Introduction to Pharmacokinetics

- Applies mathematical equations to describe drug (pharmaco) movement (kinetics) in the body following a dose
- Your body will recognize a drug as foreign (xenobiotic) and, as a defense mechanism, will immediately begin trying to eliminate it
 - Based on physicochemical properties of the xenobiotic, it can penetrate various tissues, into cells (lipophilicity, affinity for cellular transporters, etc)
- Several kinetic rates to estimate
 - Absorption
 - Distribution
 - Metabolism
 - Elimination
- While xenobiotic is being absorbed (if given non-IV) and distributed, your body is constantly working to eliminate it, either through metabolism to detoxify or make more polar, directed into intestines, or filtered by kidneys into the bladder

Pharmacokinetics (ADME) flow chart

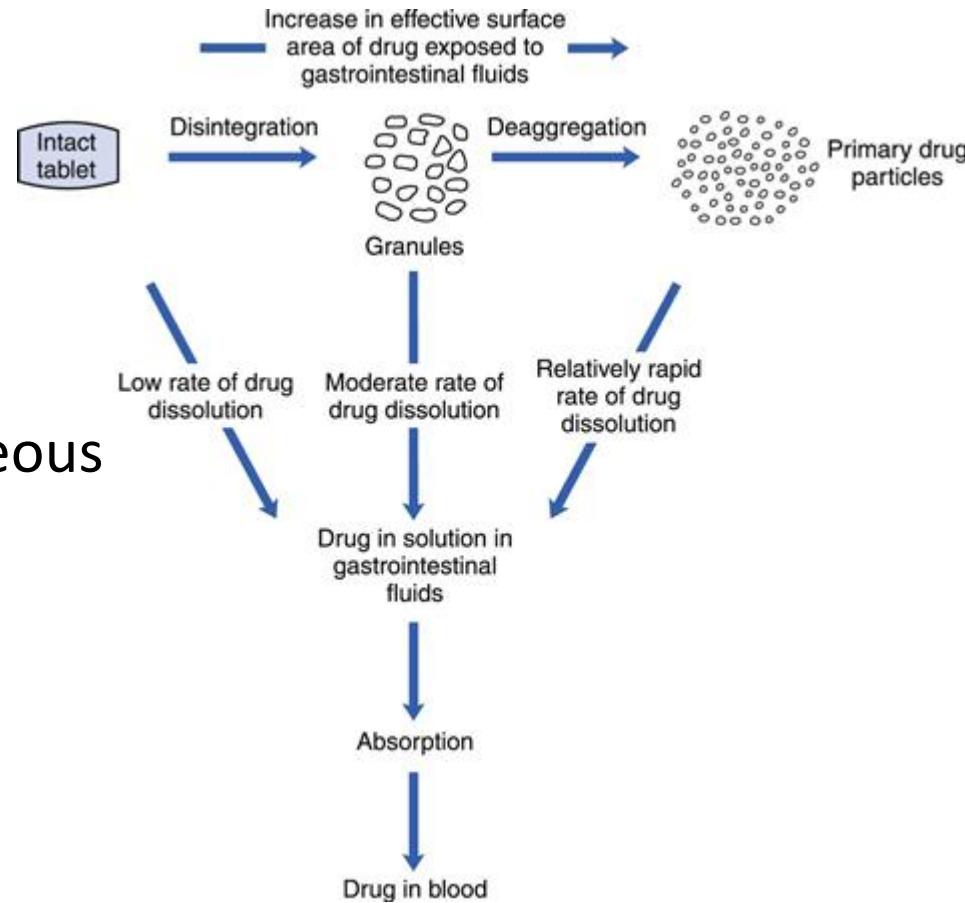


Absorption

- Applies when drug **not** given intravenously (IV)
 - IV introduces 100% of dose immediately into circulating blood flow (no absorption)
 - Where measurements of drug concentration are made
- Subcutaneous, intramuscular and other parenteral routes will have some absorption factor
 - Must pass through whatever tissue delivered in to get into blood stream
- Enteral dosing (oral, sublingual) must go through gastric and upper intestinal tract before being directed to liver
 - Following liver filtration (metabolism and/or transport via bile to lower intestines), then drug delivered to systemic circulation
 - Several rates to factor in, but normally calculate a single absorption rate for simplicity
 - Many oral dosage formulations (immediate release, extended release, tablets, capsules, solutions, suspensions, etc)
 - Each formulation has its own set of gastric dissolution rates
 - Prandial status affects gastric pH and emptying into upper intestinal tract

Tablet Absorption

- Once a tablet ingested, begins process of disintegration & deaggregation
 - Increases surface area exposed to acidic gastric aqueous fluid
- Dissolution (dissolving of drug in gastric and/or intestinal fluid) depends on the drug
 - Extent of aqueous solubility
 - pKa of the drug (weak acid or base)
 - Best chance to be dissolved if unionized
 - Depends on pH of GI fluid
- Fasted vs Fed matters
 - Alters the pH of GI fluid
 - Fed state induces bile acids and other emulsifiers to solubilize lipophilic molecules
 - Co-administered proton pump inhibitors can alter pH

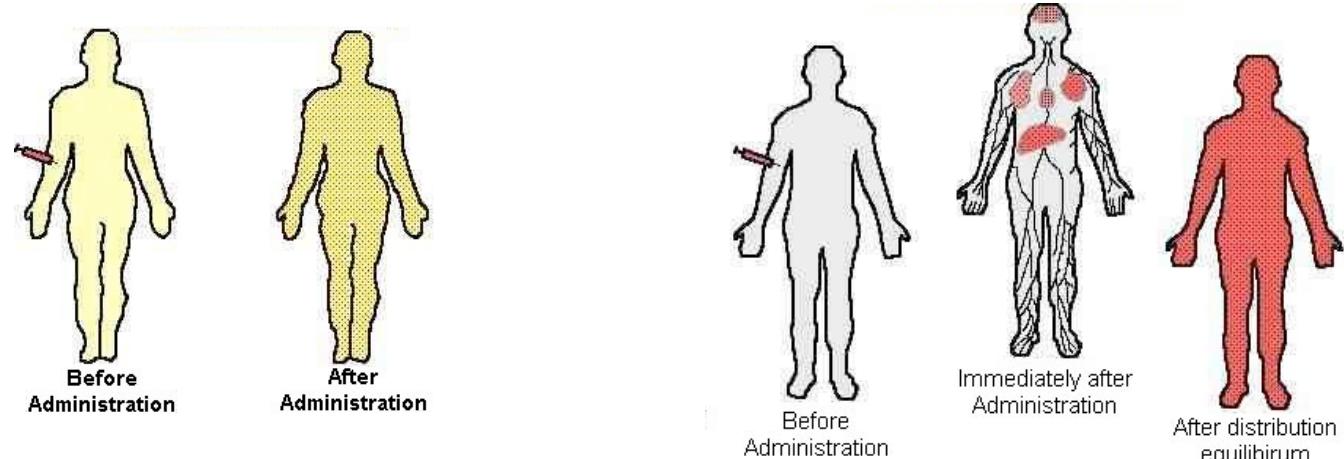


Distribution

- Two types:

1. Monophasic

- Drug distributes homogeneously
- Can assume plasma concentrations (readily measured) reflect tissue/target concentrations
- PK terms this as the drug distributing into a single “compartment”
 - Compartment is theoretical, not meant to describe an anatomical space
 - Extent of drug distribution is described by calculating the fluid **volume** of this theoretical compartment (V_1 , V_C , or V_D)



2. Polyphasic

- Drug distributes heterogeneously (based on blood flow, physicochemical drug properties etc)
- After equilibration, drug now homogeneously distributed
- Differing kinetic rates to describe drug distribution into generalized “compartments”
 - A central compartment, representing rapidly distributing areas (blood, highly-perfused organs)
 - Described by a central volume of distribution (V_C)
 - Peripheral compartment(s) represent(s) slower distributing areas (adipose tissue, bone, hair, skin, etc)
 - Described by a peripheral volume of distribution (V_p), sometimes referred to as tissue distribution volume (V_t)

Distribution

Other factors affecting drug distribution:

1. Protein binding

- Typically by human serum albumin (HSA) and α 1-acid glycoprotein (AAG)
- Only unbound drug (aka free fraction) in plasma/blood can distribute into tissues/cells
 - Distribution into kidneys leads to filtration, excretion via urination
 - Distribution into liver leads (potentially) to metabolism and transport into bile for delivery to intestines for excretion
- Presence of other drugs competing for same albumin or AAG binding sites may alter protein binding
 - Based on abundance and affinity of drugs present
 - One drug will be displaced more than the other, leading to a greater free (unbound) fraction that can distribute into a greater volume (tissues, extracellular space, intracellular), or can be metabolized or excreted
 - Measured plasma concentration of this displaced drug will be lower
- Normal albumin levels are 3.4 – 5.4 g/dL
 - Hypoalbuminemia can result in greater distribution, faster clearance, lower plasma concentration (more unbound drug)
 - Hyperalbuminemia can result in lower distribution, slower clearance, higher plasma concentration (less unbound drug)

Distribution

2. Transport

- Membrane-bound proteins that transport molecules into (influx) or out of (efflux) cells are called transporters
 - P-glycoprotein (Pgp, aka MDR1, ABCB1) is predominant drug efflux transporter
 - Located on most cell/tissue types
 - Predominantly in hepatocytes, blood-brain barrier, blood-testes barrier, blood-placental barrier, intestinal epithelium, renal proximal tubule
- Some drugs are substrates for same transporter: potential for competition
 - Usually doesn't alter distribution since Pgp and other transporter proteins so prolific
- Some drugs are inhibitors of transporters
 - This could reduce efflux to an extent, but usually not clinically relevant
 - Tariquidar, a Pgp inhibitor, never quite demonstrated significant clinical *in vivo* effect since Pgp so abundant, and due to Pgp being highly polymorphic
- Pgp, ABCC1 are two of the most studied polymorphic transporters
 - Some mutations cause a decrease or increase in expression
 - Some mutations are nonsynonymous SNPs (change in amino acid) and can cause altered folded conformation
 - Either more or less functional

The Concept of Drug Clearance

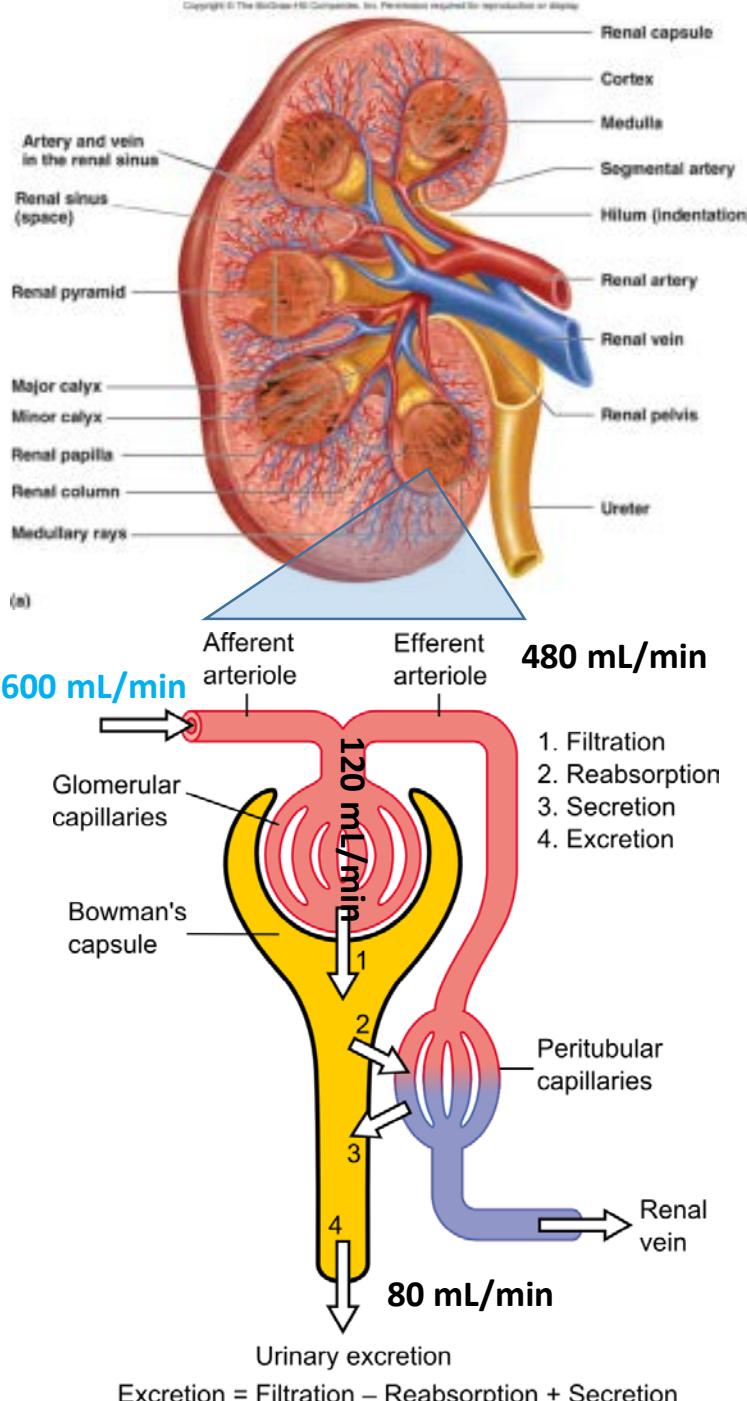
- The volume of plasma completely cleared of drug over time
 - Units of volume per time (normally L/hr or mL/min)
 - Can calculate an excretion rate (mg drug per time) by $CL * \text{concentration}$
 - $\text{mL/min} * \text{ng/mL} = \text{ng/min}$, convert to mg/hr
- $CL_T = CL_{NR} + CL_R$
 CL_R = renal clearance
 - Glomerular filtration
 - Active tubular secretion

CL_{NR} = non-renal clearance (all other routes of drug clearance not related to kidneys)

- Hepatic (CL_H)
- Catabolic
- Respiratory
- Salivary
- Biliary
 - Fecal

Renal Clearance

- Xenobiotic and any circulating metabolites in blood are filtered by kidneys
- Clearance from plasma via kidneys (CL_R) can occur:
 1. Glomerular filtration from nephron
 - Only small molecules (MW <2000 Da) that are not protein bound (free fraction)
 2. Tubular secretion from peritubular capillary
 - Active transport (ATP-dependent) to pump against concentration gradient
 - Uses transporters, hence can be saturated in multi-drug therapy, source of drug-drug interaction
- Kidneys receive ~25% of cardiac output (**600 mL plasma/min**)
 - ~20% of that flow is filtered at the glomerulus (**~120 mL/min**) = GFR
- Rate of CL_R affected by kidney function
 - Most commonly measured by creatinine clearance (CL_{CR})
 - Creatinine is a breakdown product of muscle, exclusively eliminated from body by glomerular filtration ($CL_T = CL_R$, where CL_R =filtration)
 - Cockcroft-Gault eq: $CL_{CR} = (140\text{-age}) * (\text{BW in kg}) / 72 * \text{serum creatinine}$
 - Females multiply by 0.85 (*typically* have 15% lower muscle mass)
 - Rarely directly measured (involves collecting all urine dispensed over 4-5 half-lives of the drug)

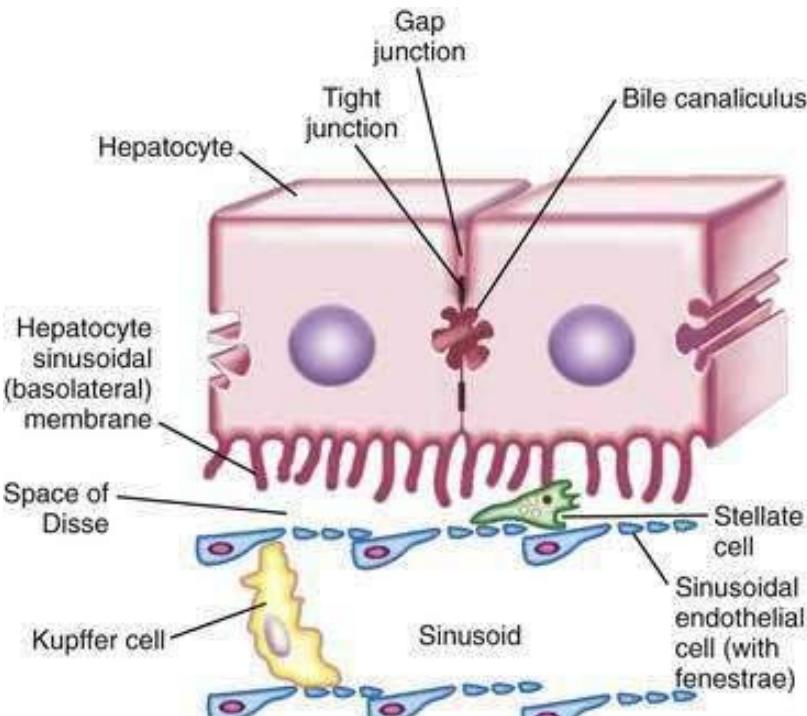


Renal Clearance (cont)

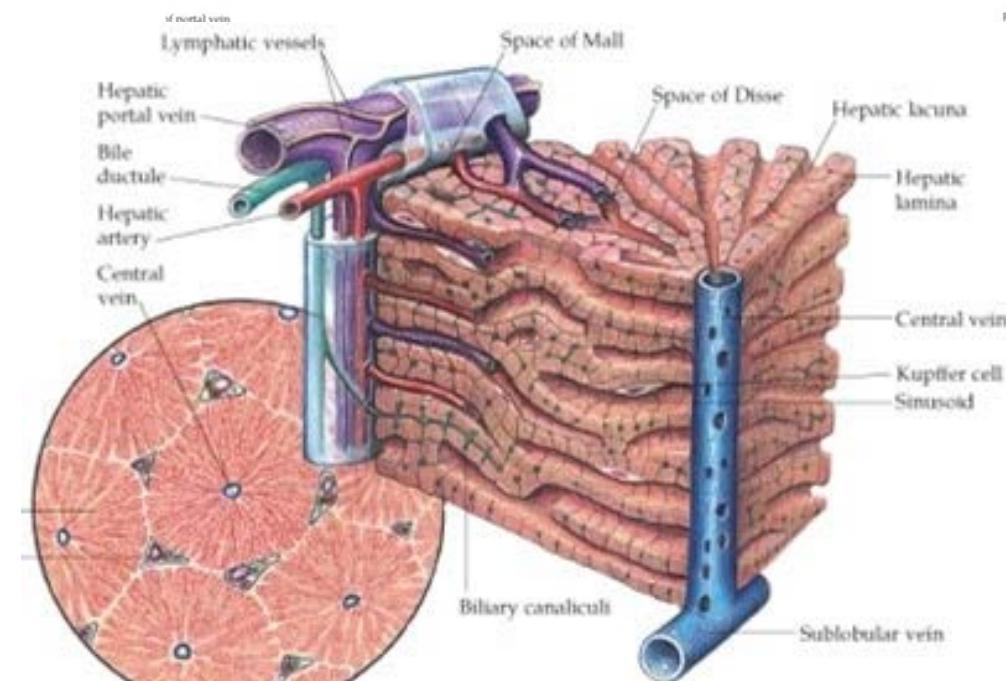
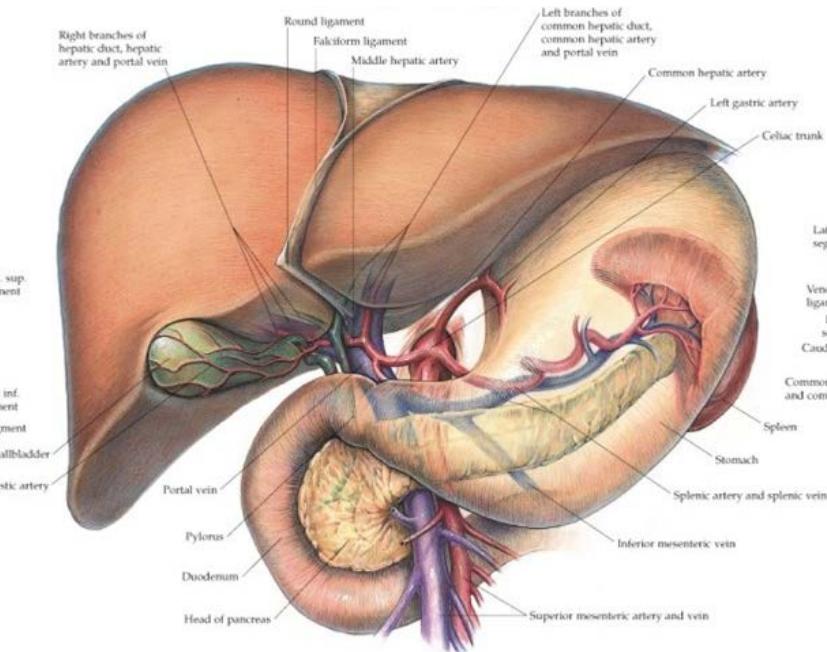
- PK studies need to be studied in renally impaired populations
 - Normal renal fnc: GFR > 90 mL/min
 - Mild renal impairment: GFR = 60 – 89 mL/min
 - Moderate renal impairment: GFR = 30– 59 mL/min
 - Severe renal impairment: GFR = 15 – 29 mL/min
 - End Stage Renal Disease: GFR < 15 mL/min (requires dialysis)
- If patients on dialysis and given a drug that is renally eliminated (even if only partially, $\sim >20\%$), dialysis will remove drug from circulation
 - Could result in subtherapeutic exposure
 - Pts on dialysis may need dose increases
- PK studies need conducted on dialysis patients (measure afferent & efferent drug conc)
 - Assess fraction of drug being removed by dialysis (C_{eff}/C_{aff})

Hepatic Clearance

- Includes hepatic metabolism
 - Largely enzymatic
- Enzymes predominantly located in smooth ER in hepatocytes
 - Transports (active and/or passive) into bile canalicula
 - Mixes with bile from the gall bladder
 - Empties in intestines for elimination
 - CL_H affected by organ function (assessed by LFTs)

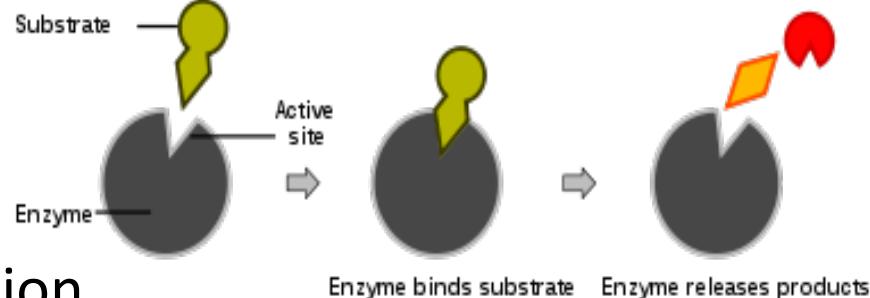


http://clinicalgate.com/wp-content/uploads/2015/05/B978141606189200072X_f11.jpg



<https://herbsforhealthandwellbeing.wordpress.com/tag/physiology-of-the-liver/>

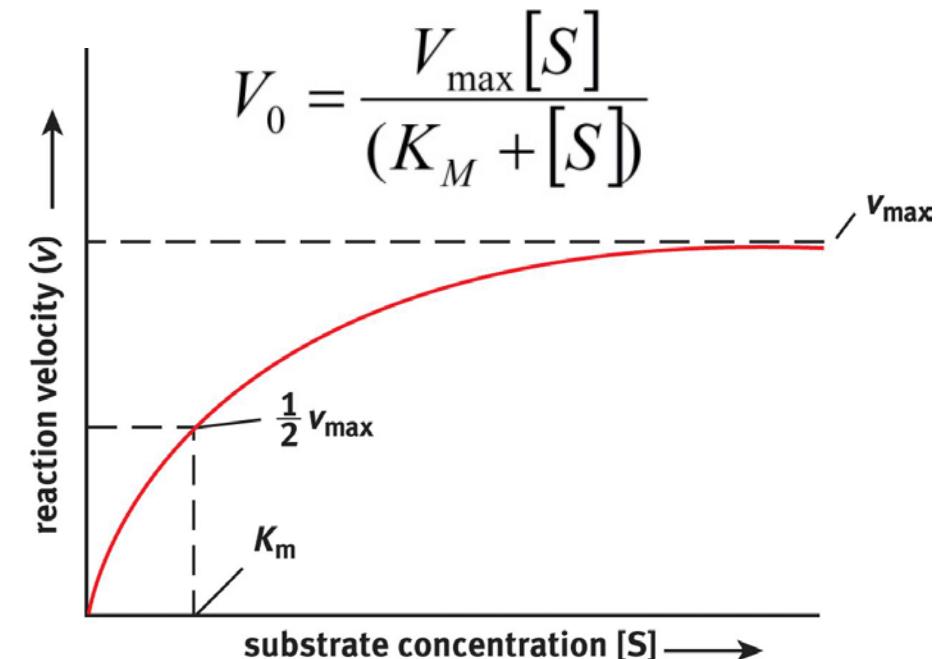
Drug Metabolism



- Many (not all) xenobiotics undergo metabolic transformation
 - Makes compound more polar to be excreted in aqueous (polar) blood, bile and urine
- Metabolism constitutes both anabolism (conjugation) and catabolism (breakdown)
- Metabolism can be spontaneous (catalyzed by pH, etc) or **enzymatically-catalyzed**:
 - Ligand (xenobiotic) binds to a protein receptor (enzyme) that catalyzes a chemical reaction (adding a polar group, removing a nonpolar group, etc)
 - Drug metabolizing enzymes predominantly found in liver and to a lesser extent in intestinal epithelia, kidneys, and some other tissues
- Major enzyme classes are cytochrome P450s (CYPs), uridine glucuronyltransferase (UGTs), Flavin monooxygenases (FMOs), glutathione transferases (GSTs), sulfotransferases, aminotransferases
- Typical reaction: $E + S \leftrightarrow ES \rightarrow P + E$
 - Read as enzyme (E) binds substrate/drug (S) to form a non-covalent complex (ES), which positions the S such that a molecular change occurs by the E to form the product (P), which is then released by the E
 - Some enzymes have a large, relatively promiscuous binding site such that many different drugs bind and are metabolized. Other enzymes have binding sites that only certain drugs can bind.

Drug Metabolism

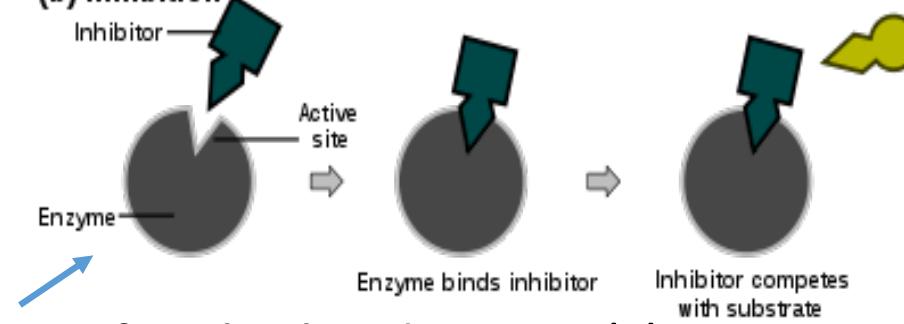
- The speed, aka velocity, of the enzyme-catalyzed reaction of metabolizing a substrate into a product is saturable
 - Only so many binding sites on so many enzymes expressed in the body
 - If more drug molecules than available binding sites, then clearance of drug via metabolism (aka intrinsic clearance) levels off (V_{max} reached) and drug molecules accumulate
 - Can cause greater than dose proportional increases in plasma exposure, leading to potential toxicity
- The substrate concentration at $\frac{1}{2}$ the maximum velocity of the reaction (V_{MAX}) is the Michaelis constant (K_M)
 - A measure of affinity of S towards a particular E
 - If two different drugs use same E, the drug with the lower K_M will preferentially be metabolized, while other drug won't, and will accumulate in blood



Changes in Drug Metabolism

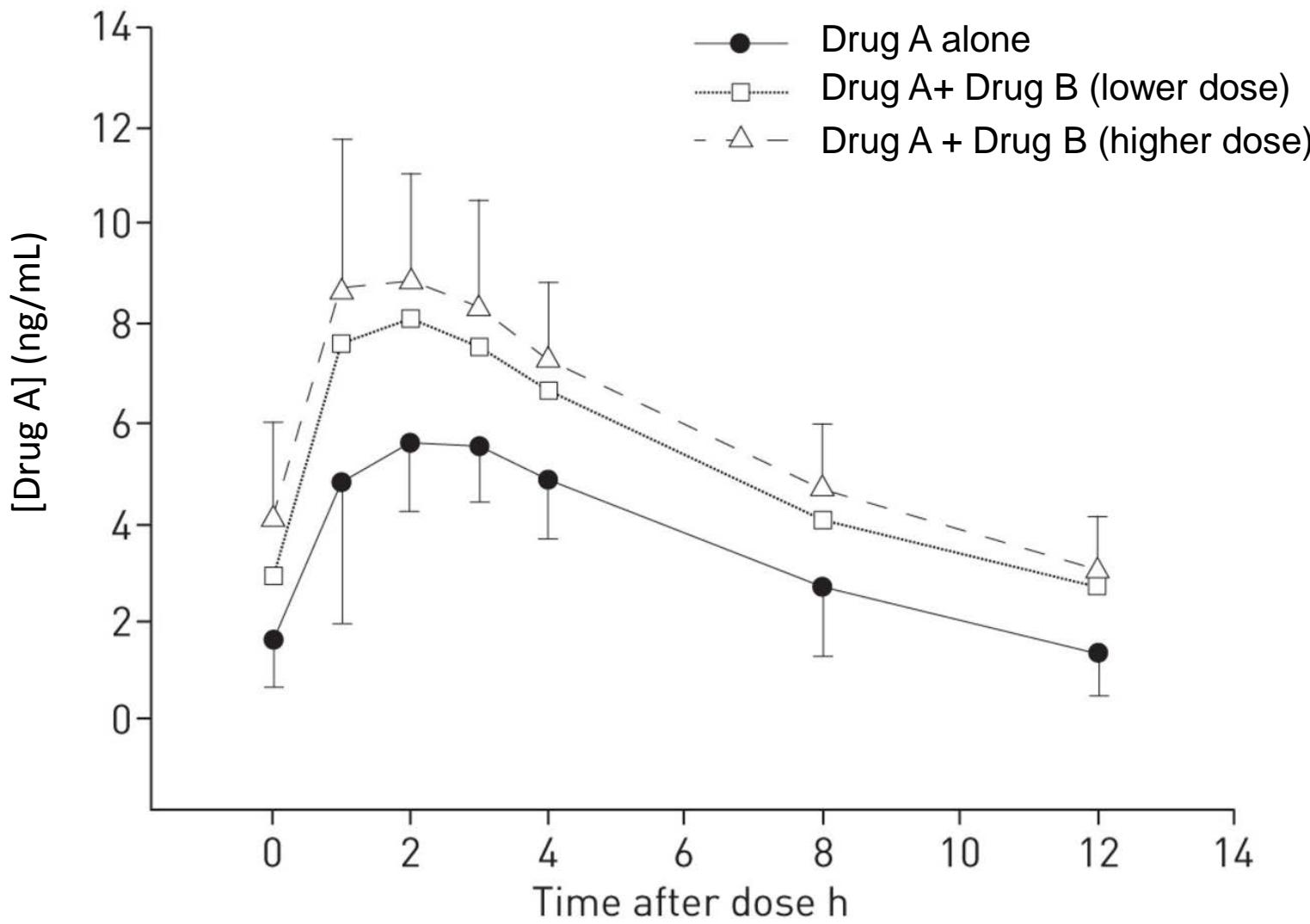
- Enzymes are proteins encoded by genes susceptible to polymorphisms
 - Can result in overexpression or vice versa, increased or decreased enzyme activity
 - An increasingly studied aspect of pharmacology
 - Optimize genotype-specific dose to avoid subtherapeutic or toxic exposure of drug
- Examples of enzymes with clinically-relevant polymorphisms:
 - CYP2C9 (*1 is wild-type, normal expression/activity, but *2 and *3 reduce metabolic activity)
 - Patients with CYP2C9*2 or *3 (either 1 or 2 copies) may need dose reductions of that drug. CYP2C9 has at least 8 other polymorphisms that are known to inactivate or decrease the function
 - CYP2D6
 - Tamoxifen is a prodrug, activated by CYP2D6-mediated metabolism into endoxifen. Patients with CYP2D6*2 or *3 will not produce as much active endoxifen as patients who are WT for CYP2D6, i.e. may need dose increase.
 - UGT1A1 (*28 confers decreased expression)
 - Patients with 1 or 2 copies of *28 taking irinotecan have greater SN38 levels b/c they can't metabolize SN38 to SN38-G as efficiently. Therefore, susceptible to SN38-induced toxicity

Inhibition of Drug Metabolism



- If two drugs (A, B) can **non-covalently** bind an E, they *compete* for the binding site(s)
 - If studying PK of drug A, then drug B would be considered a *competitive* inhibitor
 - Which ever drug had the greatest binding affinity for binding site (i.e. lower K_M), that drug would preferentially get metabolized
 - If drug A has the weaker affinity, this lack in metabolism of drug A can be overcome by overwhelming the enzyme system with drug A (will outcompete drug B based on abundance alone)
- If two drugs (C, D) can **non-covalently** bind an E, but drug D binds either to a different site on enzyme, or when it binds the substrate binding site, it alters the enzyme protein conformation so that drug C cannot bind as effectively
 - Drugs C and D are **not competing** for the same site, thus if studying PK of drug C, then drug D would be classified as a **noncompetitive** inhibitor
- If a drug binds **covalently** to enzyme, it is not reversible, hence E cannot be metabolized then released to allow enzyme to bind another drug molecule
 - This is termed an irreversible (uncompetitive) inhibition

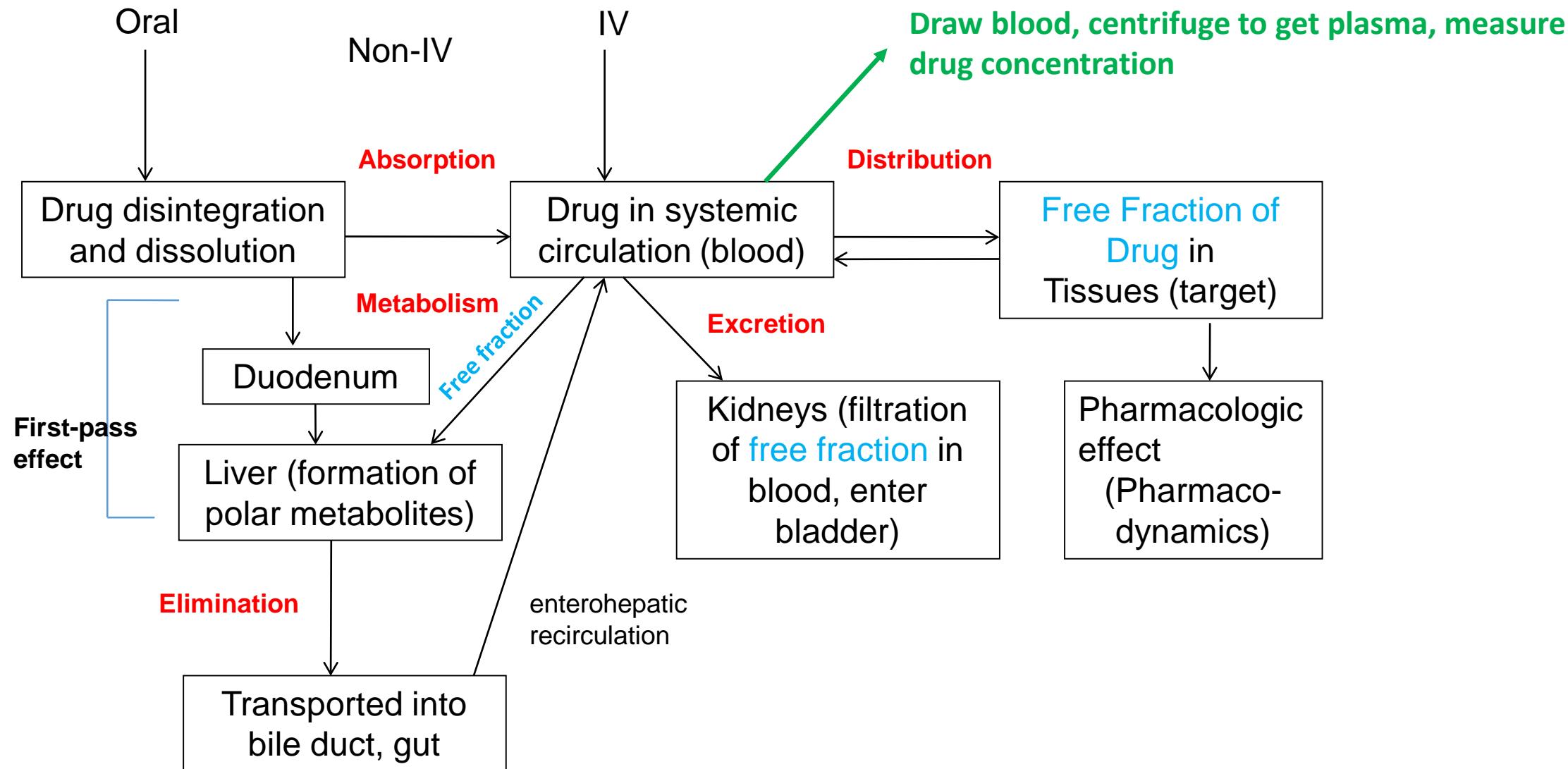
PK of Drug-Drug Interactions



How does one perform PK?

- Dose patient
- Collect blood samples at pre-specified times following dose admin (based on drug's prior estimates of half-life)
- **Measure drug concentration** in plasma at each time point
 - Plot the drug concentration vs time curve
 - Use algebra (noncompartmental) or differential calculus (compartmental) equations for PK parameters (AUC, $T_{1/2}$, CL)
 - Best to apply mathematical model to data in order to extrapolate disposition beyond observed time pts
 - Can then simulate (extrapolate) varying dose, regimens

Pharmacokinetics (ADME) flow chart



How are drug concentrations in plasma measured?

- HPLC to **separate** compounds being analyzed (analytes)
- Tandem mass spectrometry (MS/MS) to **identify, detect, and quantitate** analytes



Assay Validation

- PK plays major role in drug being FDA approved (dose amount and frequency largely determined by PK)
 - LC-MS/MS has major influence over PK
 - FDA regulates how LC-MS/MS assays conducted to produce most reliable measurements for most reliable PK
- FDA requires less than 15% deviation from expected conc (accuracy) and 15% precision over a 4-day period
- Efficient extraction
- Stability
 - in solution
 - in matrix
 - freeze/thaw cycles

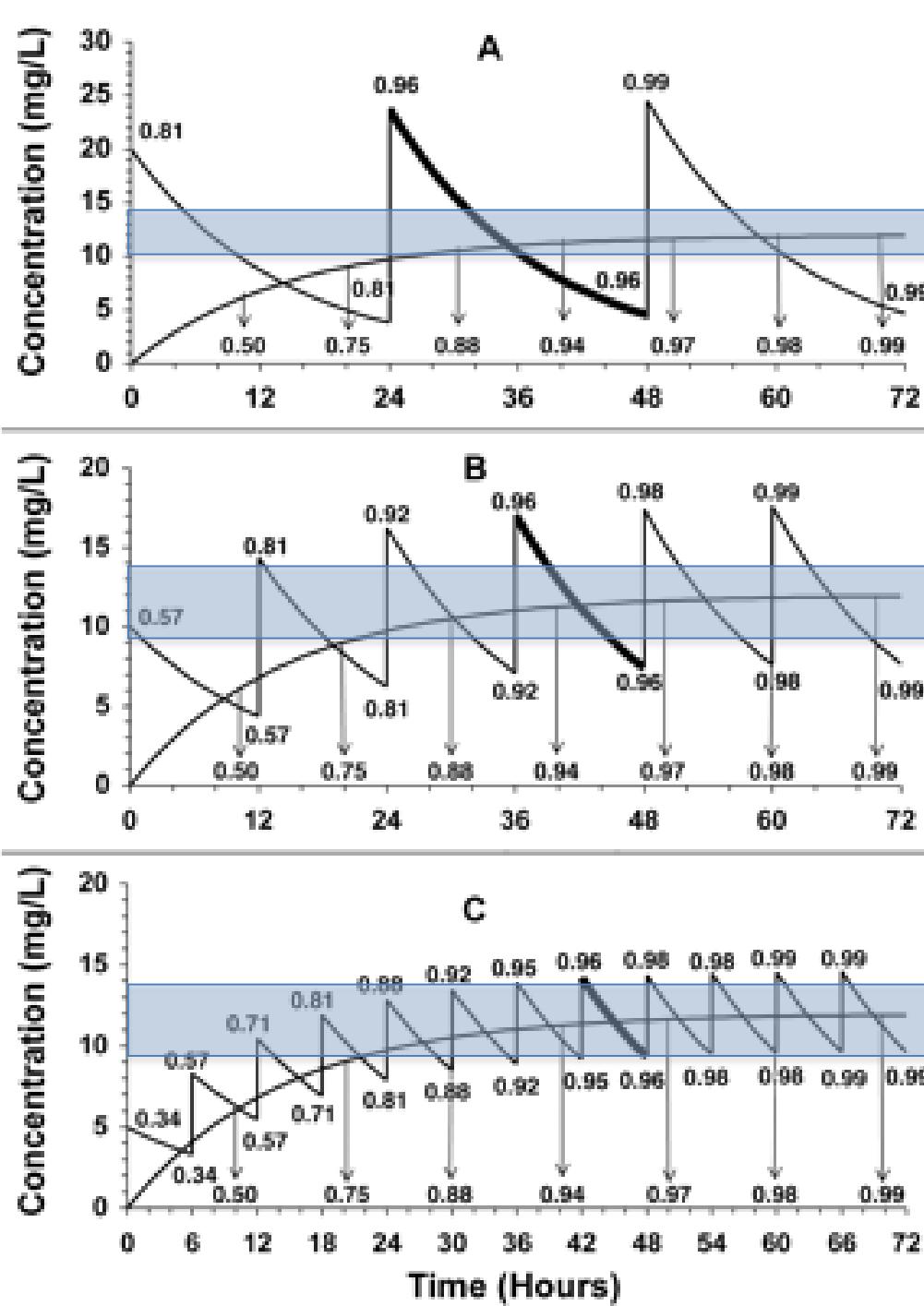
Bottom Line: **Must be able to trust data**

How does one perform PK?

- Dose patient
- Collect blood samples at specified times following dose admin
- Measure drug concentration in plasma at each time point
- Plot the drug concentration vs time curve
- Use algebra (noncompartmental) or differential calculus (compartmental) equations for PK parameters (AUC, $T_{1/2}$, CL)
 - Best to apply mathematical model to data in order to extrapolate disposition beyond observed time pts
 - Can then simulate (extrapolate) varying dose, regimens

Uses of Pharmacokinetics

- Predict plasma exposure
 - Based on drug's dose, route, tau, ADME
- Predict toxicities based on exposure
 - Assuming linear (dose-proportional) PK
- Identify drug-drug interactions
- Provide optimized dose schedule for most efficacious/least toxic therapy



Tau=24 hrs

Tau=12 hrs

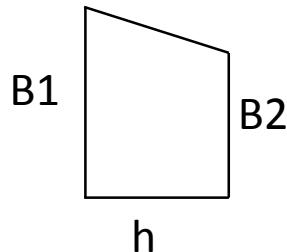
Tau=10 hrs

Noncompartmental Analysis (NCA)

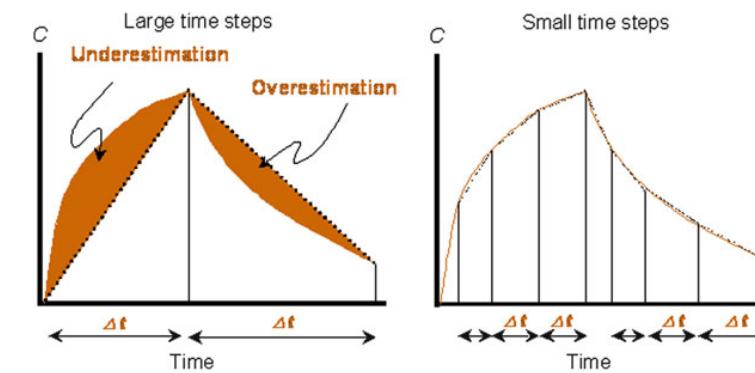
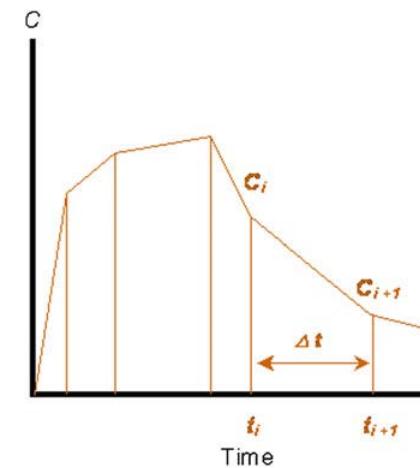
- Not a mathematical “model”
- Uses integral calculus to calculate AUC from trapezoidal rule
 - With AUC (and knowing the dose), can calculate CL, and Volume
 - Cmax, Tmax are observed values
- Simple, therefore commonly used for preclinical and early-stage clinical trials
- This workshop will only use PK data calculated using NCA
 - No compartmental or population analyses used here

Trapezoidal Rule

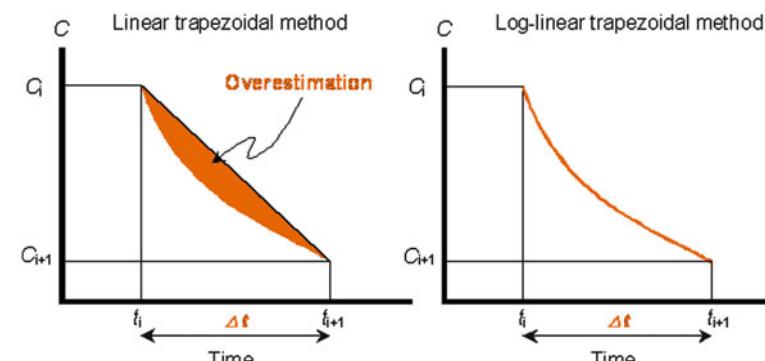
- Uses log-linear trapezoidal rule to calculate AUC_{ALL}
 - Area of trapezoid = $\frac{(\text{Base1} + \text{Base2}) * \text{height}}{2}$



- Linear trapezoid over/underestimates
 - decrease step size
- Linear also overestimates (regardless of step size) descending curves of first-order processes (most)
 - need exponential decay
- Log-Linear best: uses linear on ascending and flat and log on descending



$$AUC_0^{t_{\text{last}}} = \sum_{i=1}^n \frac{C_i + C_{i+1}}{2} \cdot \Delta t, \quad AUC_0^{t_n} = \sum_{i=1}^n \frac{C_i - C_{i+1}}{\ln(C_i/C_{i+1})} \cdot \Delta t,$$



NCA Equations

- Calculate elimination rate (λ_z)

- Can calculate AUC_{EXTRAP} ($AUC_{t_{\text{last}}}^{\infty} (\text{observed}) = \frac{C_{\text{last}}}{\lambda_z}$)

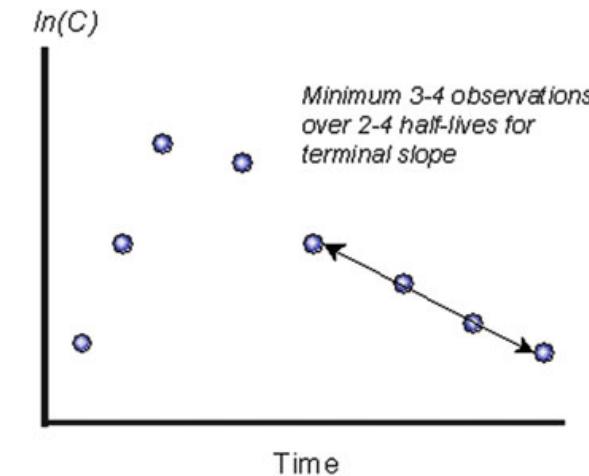
- $AUC_{\infty} = AUC_{\text{ALL}} + AUC_{\text{EXTRAP}}$

- Half-Life: $t_{1/2z} = \frac{\ln(2)}{\lambda_z}$

- Mean Residence Time: $MRT = \frac{AUMC_0^{\infty}}{AUC_0^{\infty}}$ $AUMC_0^{t_{\text{last}}} = \sum_{i=1}^n \frac{t_i \cdot C_i + t_{i+1} \cdot C_{i+1}}{2} \cdot \Delta t.$

- **CLEARANCE:** $Cl = \frac{D_{\text{iv}}}{AUC_0^{\infty}}$ $Cl_0 = \frac{Cl}{F} = \frac{D_{\text{po}}}{AUC_0^{\infty}}$. $F = \frac{AUC_{\text{ev}}}{AUC_{\text{iv}}} \cdot \frac{D_{\text{iv}}}{D_{\text{ev}}},$

- **VOLUME of DISTRIBUTION:** $V_{\text{ss}} = MRT \cdot Cl$



Advantages of NCA

- Relatively simple to perform
- Do not need expertise in PK modeling and simulation
- Each subject/patient would have their own parameter values
- Can average the parameter values for all subjects in a subgroup, cohort, etc.
- Fast and easy way to identify if there are any major differences in a PK parameter
- Can use these exposure metrics to correlate with pharmacodynamic response data

Lunch

NCA using R

- NCA often conducted using WinNonlin®
 - These require expensive annual licenses
 - Can also use Matlab and Microsoft Excel®
 - These cost \$ as well, and aren't designed for PK analyses
 - Requires training and understanding of NCA equations for user to enter manually
- R is free
- Anyone can write a package for use in R
- Install packages, then load them
- Need basic coding skills

NCA using R

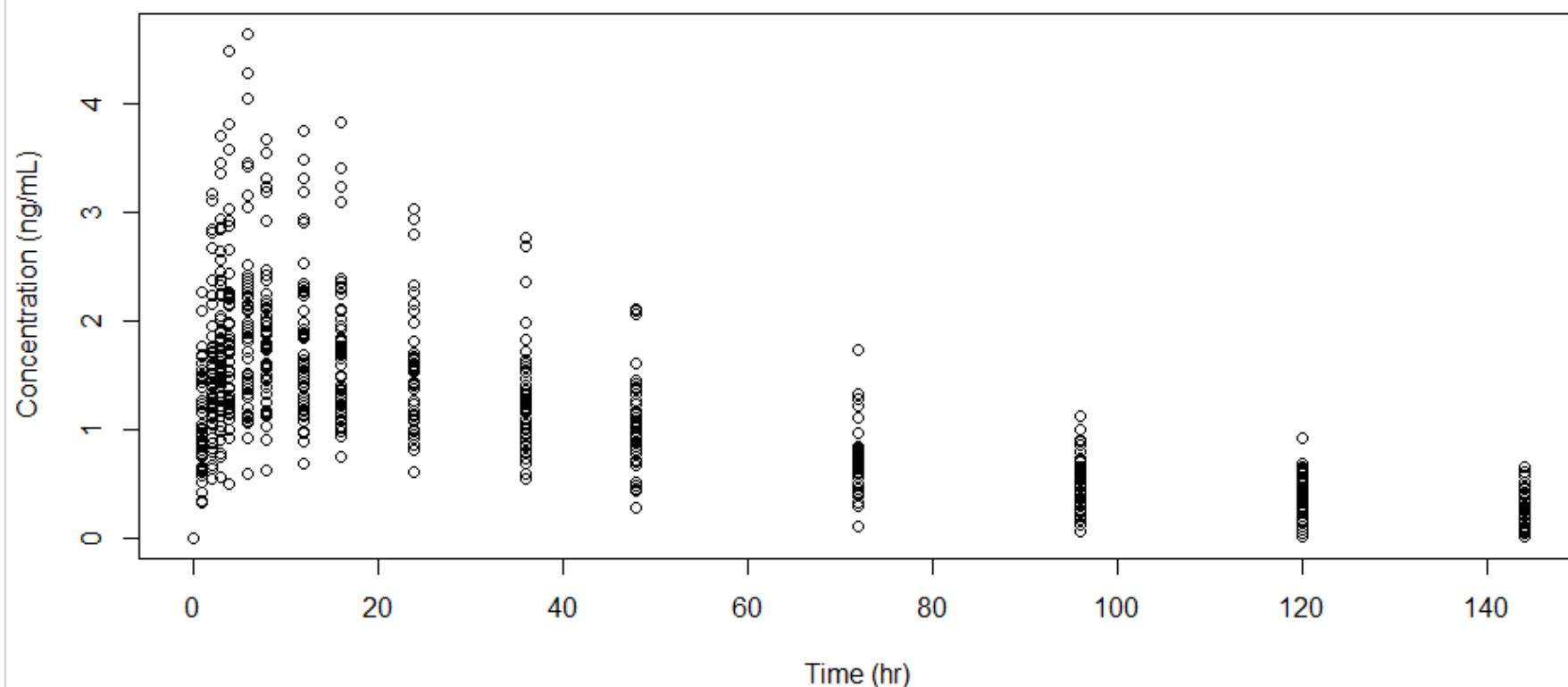
- Use “finalPKdata.csv” dataset for a blood pressure drug and blood pressure clinical trial
- Read in file: `data1<-read.csv("finalPKdata.csv", header=TRUE)`
- Install “nlme”, “PKNCA”, “knitr”, and “ggplot2” packages
 - `install.packages("nlme")`
- Activate required library packages

```
#### activate required libraries ####  
library(PKNCA)  
library(ggplot2)
```

Plot Conc vs Time Data

- Many ways to make XY plots
 - Can simply use built-in “plot” function:

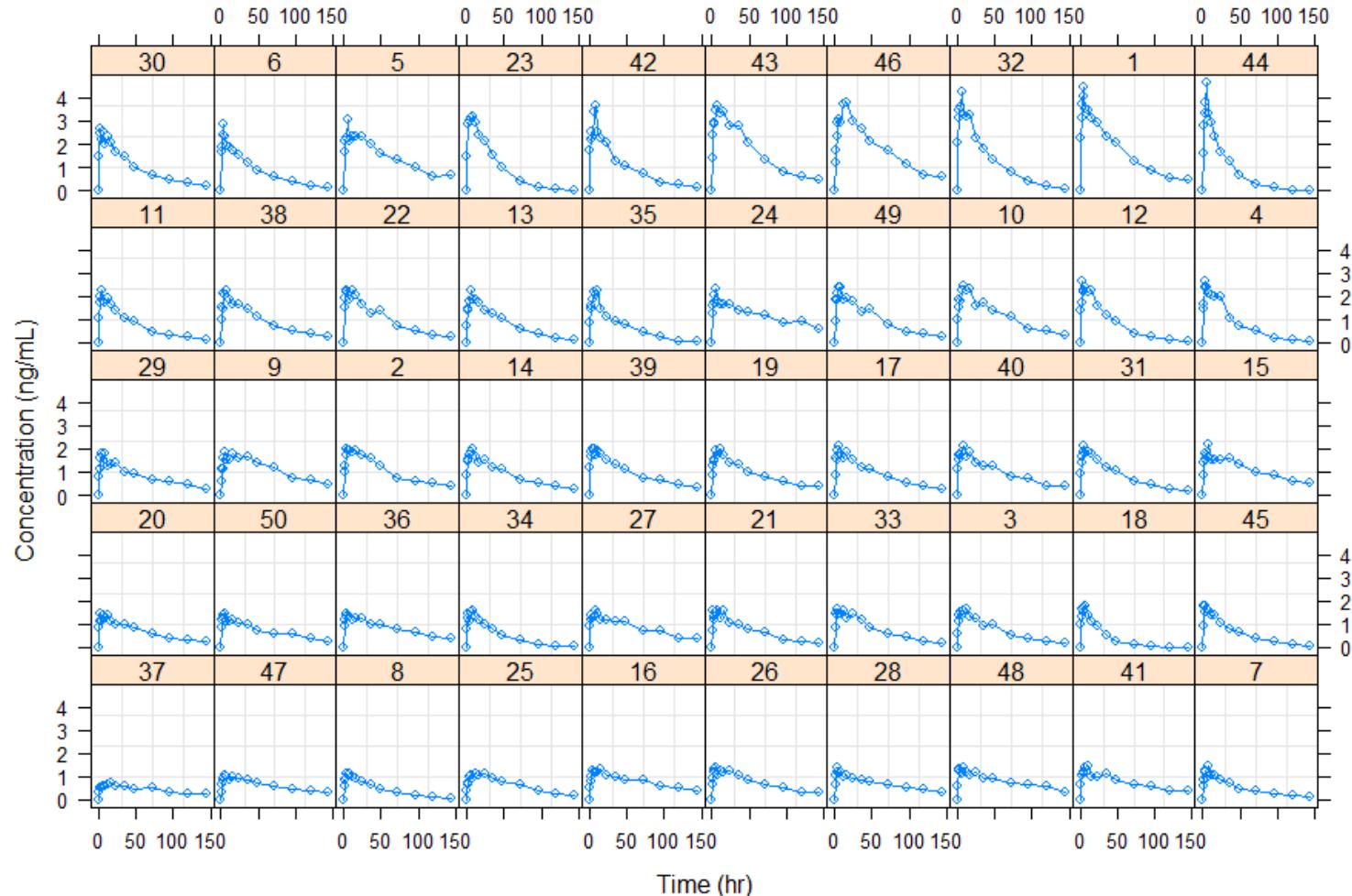
```
plot(data1$conc~data1$time, ylab="Concentration (ng/mL)", xlab="Time (hr)")
```



Separate Plots by Subject

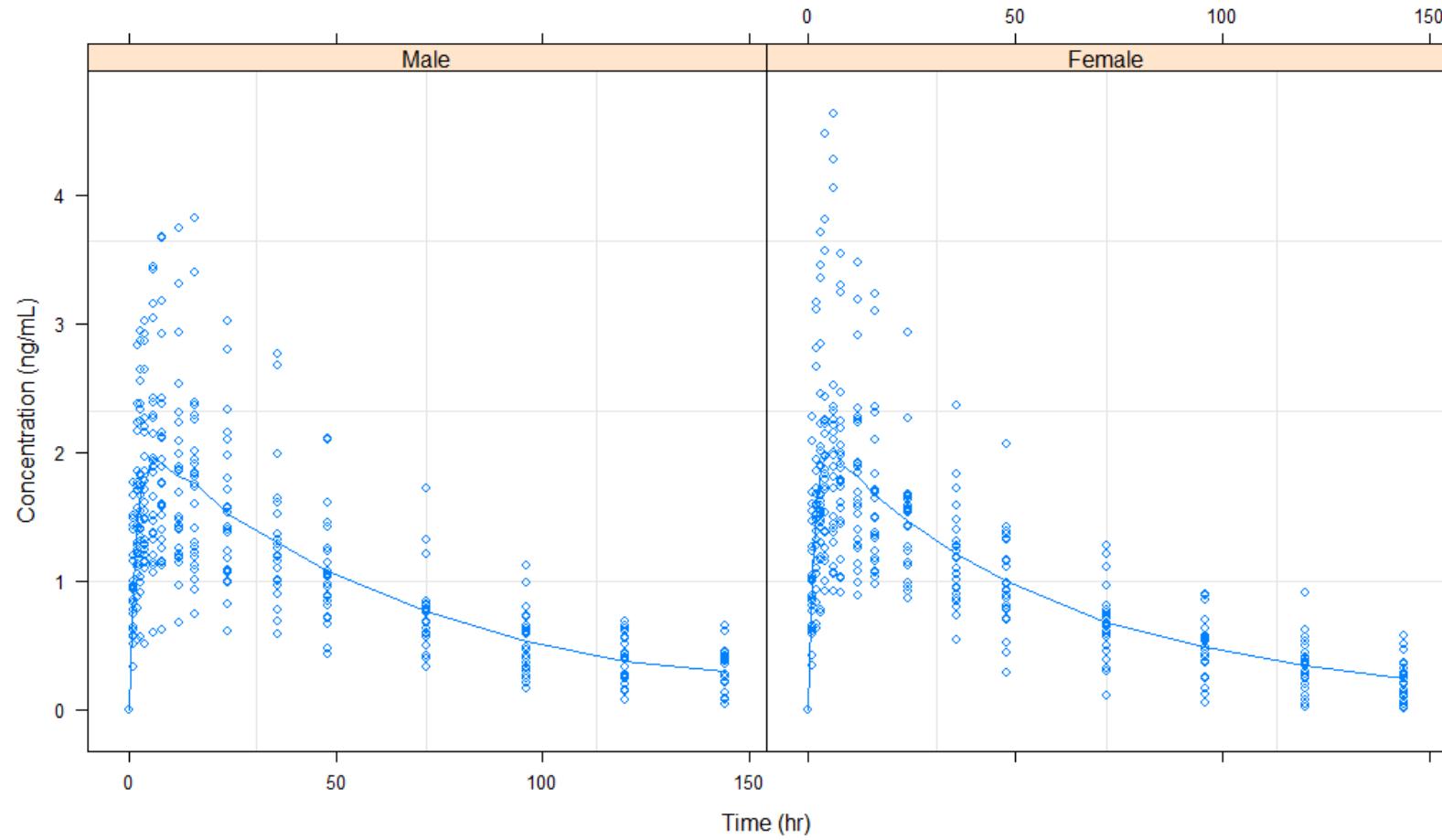
```
#### group data and plot ####
data1.2<-groupedData(conc~Time|ID, data=data1, labels = list(x="Time(hr)", y="Conc(ng/mL)"))

plot(data1.2, aspect=1/1, ylab="Concentration (ng/mL)", xlab="Time (hr)")
```



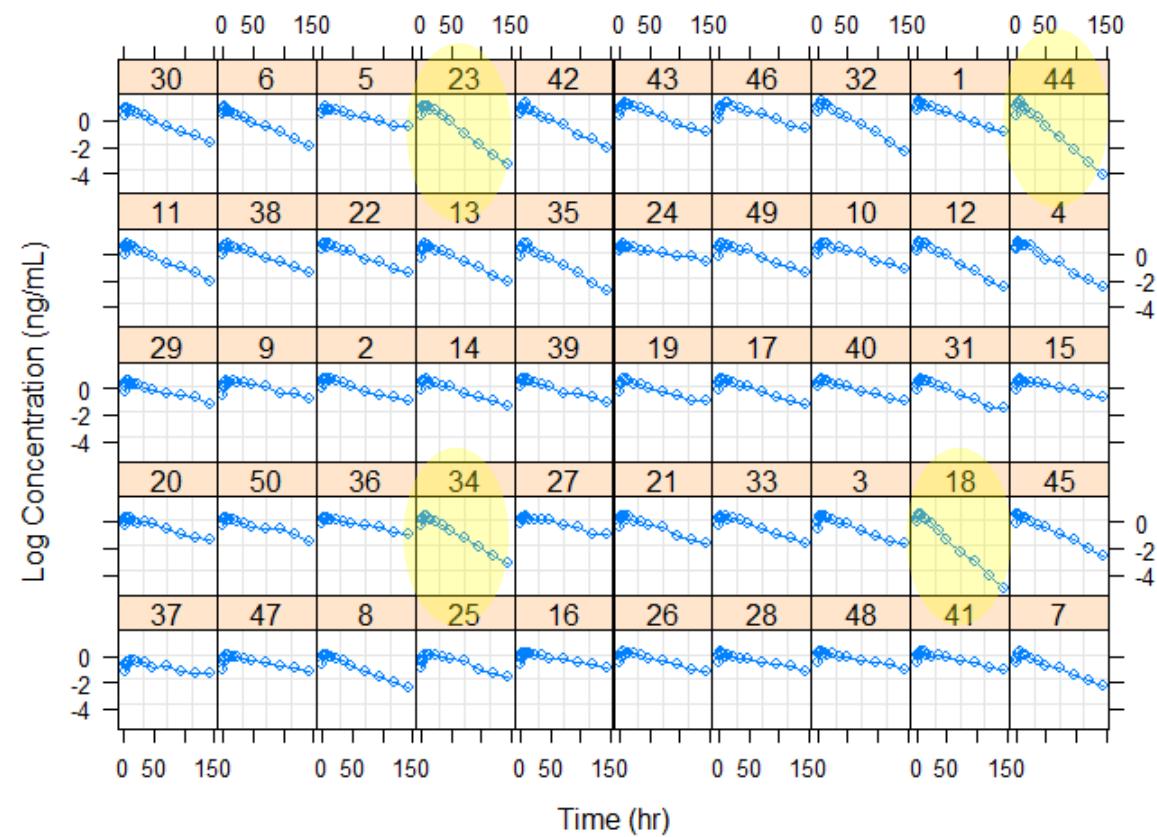
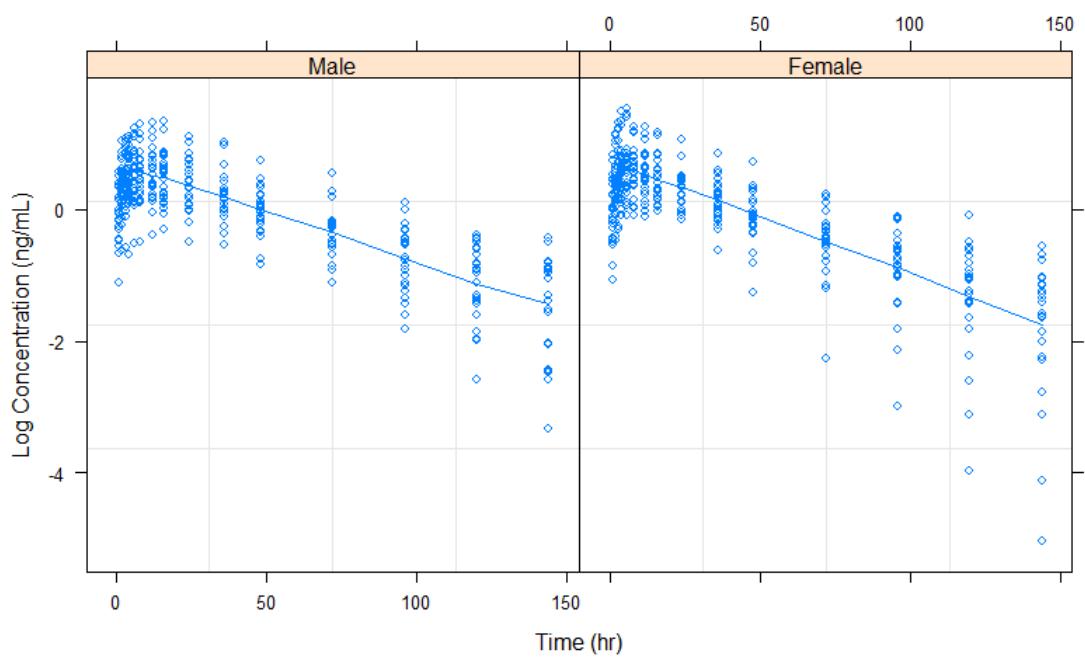
Plot by Biological Sex

```
#### PK Profiles by Sex ####  
data1.3<-groupedData(Conc~Time|Gender, data=data1, labels=list(x="Time", y="Concentration"))  
plot(data1.3, aspect=1/1, ylab="Concentration (ng/mL)", xlab="Time (hr)")
```



Log Transform Data

```
### read in data ###
data1<-read.csv("data1.csv", header=TRUE)
data1$logConc<-log(data1$Conc)
```



Perform NCA

- Now that we've plotted data, there appears to be no outliers
 - Very normal-looking PK profiles following oral administration (absorption, mono-exponential elimination)

```
## By default it is groupedData; convert it to a data frame for use--  
### maps conc to any variable or grouping factor ##  
my.conc <- PKNCAConc(as.data.frame(data1), Conc~Time|ID)  
  
## Dosing data needs to only have one row per dose, so subset for that first##  
d.dose <- unique(data1[data1$Time == 0,  
                           c("ID", "Time", "Dose", "Conc", "Age", "Weight", "SCreatinine", "SerumALT", "Gender", "Race")])  
knitr::kable(d.dose,  
             caption="Dosing data extracted from data set")
```

```
### by dose  
my.dose <- PKNCAdose(d.dose, Dose~Time|ID)  
my.dose  
#### combine dose and conc data ##  
my.data.automatic <- PKNCAdata(my.conc, my.dose)  
summary(my.data.automatic)  
knitr::kable(PKNCAdoptions("single.dose.aucs"))  
knitr::kable(my.data.automatic$intervals)
```

```
### compute parameters ###  
my.results.automatic <- pk.nca(my.data.automatic)  
summary(my.results.automatic)
```

Automatic/default settings don't include clearance or volume of distribution

Parameter	Estimate (%CV)
Cmax (mg/L)	2.02 (40.5)
Tmax (hr)	7.0 (range 2-24hr)
HL (hr)	50.3 (20.2%)
AUC _{0-24hr} (hr*mg/L)	37.8 (36.8%)
AUC _{INF} (hr*mg/L)	137 (36.9%)

Perform NCA

- Notice how much larger AUC_{INF} is vs $AUC_{0-24\text{hr}}$
- Half-life is 50.3 hr
 - A slow-eliminating drug
- A 24-hr time interval for capturing AUC is not very informative
- Could do 1 of two things...
 1. Only focus on AUC_{INF}
 - Reasonable b/c plots showed a nice straight line that had good correlation (r^2) for estimation of elimination rate, which is required to calculate AUC_{INF}
 - More informative for a first dose PK analysis
 - Useful for comparing and predicting the AUC over a dose interval (QD, BID, QOD, etc)
 2. Lengthen time interval for AUC to 0-96hr or 0-120hr (typically 4-5 half-lives)

Parameter	Estimate (%CV)
Cmax (mg/L)	2.02 (40.5)
Tmax (hr)	7.0 (range 2-24hr)
HL (hr)	50.3 (20.2%)
$AUC_{0-24\text{hr}}$ (hr*mg/L)	37.8 (36.8%)
AUC_{INF} (hr*mg/L)	137 (36.9%)

NCA – Manual Intervals

- Manually determine which parameters estimated...

```
my.intervals <- data.frame(start=0,
                             end=Inf,
                             cmax=TRUE,
                             tmax=TRUE,
                             aucinf=TRUE,
                             auclast=FALSE,
                             cl=TRUE,
                             vz=TRUE,
                             half.life=TRUE)

my.data.manual <- PKNCAdata(my.conc, my.dose,
                             intervals=my.intervals)
knitr::kable(my.data.manual$interval)
summary(my.data.manual)

my.results.manual <- pk.nca(my.data.manual)
summary(my.results.manual)
'
summary(my.results.manual)
start end      cmax      tmax half.life    aucinf      cl      vz
 0 Inf 2.02 [40.5] 7.00 [2.00, 24.0] 50.3 [20.2] 137 [36.9] 36.5 [36.9] 2430 [41.9]
```

Parameter	Estimate (%CV)
Cmax (mg/L)	2.02 (40.5)
Tmax (hr)	7.0 (range 2-24hr)
HL (hr)	50.3 (20.2%)
AUC _{INF} (hr*mg/L)	137 (36.9%)
CL (L/hr)	36.5 (36.9%)
Vz (L)	2430 (41.9%)

NCA by Sex

- Subset out data by Sex, re-run

```
##### PK by SEX #####
my.conc.sex <- PKNCACconc(as.data.frame(data1), Conc~Time|Gender+ID)
my.dose.sex <- PKNCAdose(d.dose, Dose~Time|Gender+ID)
my.data.auto.sex <- PKNCAdata(my.conc.sex, my.dose)
my.results.auto.sex <- pk.nca(my.data.auto.sex)
summary(my.results.auto.sex)

my.data.manual.sex <- PKNCAdata(my.conc.sex, my.dose.sex,
                                    intervals=my.intervals)
my.results.manual.sex <- pk.nca(my.data.manual.sex)
summary(my.results.manual.sex)

my.results.manual.sex <- pk.nca(my.data.manual.sex)
summary(my.results.manual.sex)

  start end Gender      cmax          tmax half.life    aucinf        cl         vz
  0 Inf Female 2.08 [39.6] 6.00 [2.00, 24.0] 47.2 [19.4] 130 [38.3] 38.3 [38.3] 2390 [40.3]
  0 Inf   Male 1.96 [42.0] 8.00 [3.00, 16.0] 53.4 [20.9] 144 [35.6] 34.7 [35.6] 2470 [44.3]
```

Parameter	Estimate (%CV)	
	Males	Females
Cmax (mg/L)	1.96 (42%)	2.08 (40%)
Tmax (hr)	8 (range 3-16hr)	6 (range 2-24hr)
HL (hr)	53.4 (21%)	47.2 (19%)
AUC _{INF} (hr*mg/L)	144 (36%)	130 (38%)
CL (L/hr)	34.7 (36%)	38.3 (38%)
Vz (L)	2470 (44%)	2390 (40%)

NCA by Race

Parameter	Estimate (%CV)				
	Asian	Black	Caucasian	Hispanic	Other
Cmax (mg/L)	1.90 (24%)	1.84 (46%)	1.97 (43%)	2.71 (50%)	2.09 (21%)
Tmax (hr)	4.5 (2-16)	6 (3-24)	8 (3-16)	6 (4-8)	6 (4-12)
HL (hr)	47.5 (7.4%)	46.5 (17%)	50.5 (24%)	60.5 (27%)	50.3 (9.8%)
AUC _{INF} (hr*mg/L)	139 (25%)	122 (39%)	128 (35%)	204 (22%)	151 (37%)
CL (L/hr)	36 (25%)	41 (39%)	39 (35%)	24 (22%)	33 (37%)
Vz (L)	2440 (20%)	2590 (45%)	2510 (46%)	1920 (57%)	2360 (25%)

#####
PK by Race #####

```
my.conc.race <- PKNCACconc(as.data.frame(data1), Conc~Time|Race+ID)
my.dose.race <- PKNCAdose(d.dose, Dose~Time|Race+ID)
my.data.auto.race <- PKNCAdata(my.conc.race, my.dose)
my.results.auto.race <- pk.nca(my.data.auto.race)
summary(my.results.auto.race)
```

```
my.data.manual.race <- PKNCAdata(my.conc.race, my.dose.race,
intervals=my.intervals)
my.results.manual.race <- pk.nca(my.data.manual.race)
summary(my.results.manual.race)
```

```
summary(my.results.manual.race)
```

start	end	Race	cmax	tmax	half.life	aucinf	c1	vz
0	Inf	Hispanic	2.71 [50.1]	6.00 [4.00, 8.00]	60.5 [27.3]	204 [22.0]	24.5 [22.0]	1920 [56.9]
0	Inf	Other	2.09 [20.6]	6.00 [4.00, 12.0]	50.3 [9.76]	151 [37.0]	33.0 [37.0]	2360 [24.7]
0	Inf	Caucasian	1.97 [43.4]	8.00 [3.00, 16.0]	50.5 [24.4]	128 [35.0]	39.1 [35.0]	2510 [46.1]
6	0	Black	1.84 [45.8]	6.00 [3.00, 24.0]	46.5 [17.2]	122 [39.2]	41.1 [39.2]	2590 [45.4]
1	0	Asian	1.90 [23.7]	4.50 [2.00, 16.0]	47.5 [7.43]	139 [25.2]	36.0 [25.2]	2440 [20.3]

Obtain Individual PK Data

- R automatically provides mean, standard deviation
- Can subject-specific data

```
my.results.manual <- pk.nca(my.data.manual)
summary(my.results.manual)
my.results.manual$result
```

- Not the optimal way of presenting data...
 - Each parameter in a row
 - Would prefer the parameters be columns with 50 rows for each of the 50 subjects..

start	end	ID	PPTESTCD	PPORRES
0	Inf	1	auclast	2.309236e+02
0	Inf	1	cmax	4.479628e+00
0	Inf	1	tmax	4.000000e+00
0	Inf	1	tlast	1.440000e+02
0	Inf	1	lambda.z	1.595606e-02
0	Inf	1	r.squared	9.943818e-01
0	Inf	1	adj.r.squared	9.937576e-01
0	Inf	1	lambda.z.time.first	6.000000e+00
0	Inf	1	lambda.z.n.points	1.100000e+01
0	Inf	1	clast.pred	4.208275e-01
0	Inf	1	half.life	4.344099e+01
0	Inf	1	span.ratio	3.176723e+00
0	Inf	1	aucinf	2.600867e+02
0	Inf	1	cl	1.922436e+01
0	Inf	1	vz	1.204831e+03
0	Inf	2	auclast	1.442049e+02
0	Inf	2	cmax	1.967053e+00
0	Inf	2	tmax	4.000000e+00
0	Inf	2	tlast	1.440000e+02
				- - - - -

“Melting” and “Reshaping” Data

- Can “reshape” the data to put parameters in columns
- First need to “melt” the data using the “reshape2” package

```
install.packages("reshape2")
```

```
| library(reshape2)
```

Optimizing the Dataframe

- The subject-specific PK data is put into a data frame:

```
### make a data.frame for PK manual results###
PKresults<-my.results.manual$result
```

- Don't need the first two columns, so drop them:

```
PKresults1<-PKresults[c(-1, -2)]
```

- "Melt" the data down:

```
### melt PK data ####
mPKresults<-melt(PKresults1, id=c("ID", "PPTESTCD"))
```

- Cast, or "reshape" the data the order you want:

```
### cast the melted data ##
### cast(data, formula, function) ####
PKresults3<-cast(mPKresults, ID~variable+PPTESTCD)
```

start	end	ID	PPTESTCD	PPORRES
0	Inf	1	auclast	2.309236e+02
0	Inf	1	cmax	4.479628e+00
0	Inf	1	tmax	4.000000e+00
0	Inf	1	tlast	1.440000e+02
0	Inf	1	lambda.z	1.595606e-02
0	Inf	1	r.squared	9.943818e-01
0	Inf	1	adj.r.squared	9.937576e-01
0	Inf	1	lambda.z.time.first	6.000000e+00
0	Inf	1	lambda.z.n.points	1.100000e+01

ID	PPORRES_auclast	PPORRES_cmax	PPORRES_tmax	PPORRES_tlast	PPORRES_adj.r.square
1 1	230.92363	4.4796275		4 144	0.993757
2 2	144.20495	1.9670532		4 144	0.969844
3 3	98.96674	1.6851715		12 144	0.987549
4 4	109.42164	2.6469558		4 144	0.989162
5 5	199.19053	3.0418756		6 144	0.970430
6 6	116.25162	2.8404292		3 144	0.997388
7 7	66.46543	1.4461674		8 144	0.995562
8 8	60.49117	1.1164224		8 144	0.997896
9 9	157.67049	1.8364152		6 144	0.967593
10 10	161.55660	2.4664270		8 144	0.981246
11 11	105.78697	2.2364899		4 144	0.993740
12 12	110.39769	2.6397564		3 144	0.994740
13 13	111.41935	2.2671862		6 144	0.993785

Analyzing Relationships in Data

- PK data sometimes correlated with demographics
 - Age can hinder drug metabolism
 - Larger body weights often correlate with large volumes of distribution, etc
- Before can analyze for potential correlations, must first add demographic data columns from original dataset to the newly reshaped subject-specific PK dataset

- Original dataframe (“data1”) has several time points with drug concentrations per subject
 - For this purpose, we don’t need each time point
 - Only need demographic data columns, which is same in every row per subject
 - Thus, subset out only one row per subject (e.g. time zero)

```
#### subset out only t=0hr rows in main data.frame ####  
data0<-data1[data1$Time==0,]  
### drop the log conc column ###  
data0.1<-data0[c(-11)]
```

- Now, combine columns from “data0.1” with “PKResults3” dataframe
 - Uses the column binding function “cbind”

```
#### Cbind pk results to subsetted data.frame ####  
alldata<-cbind2(PKResults3, data0.1)
```

- The “alldata” dataframe now ready for correlative analyses

PK Correlatives

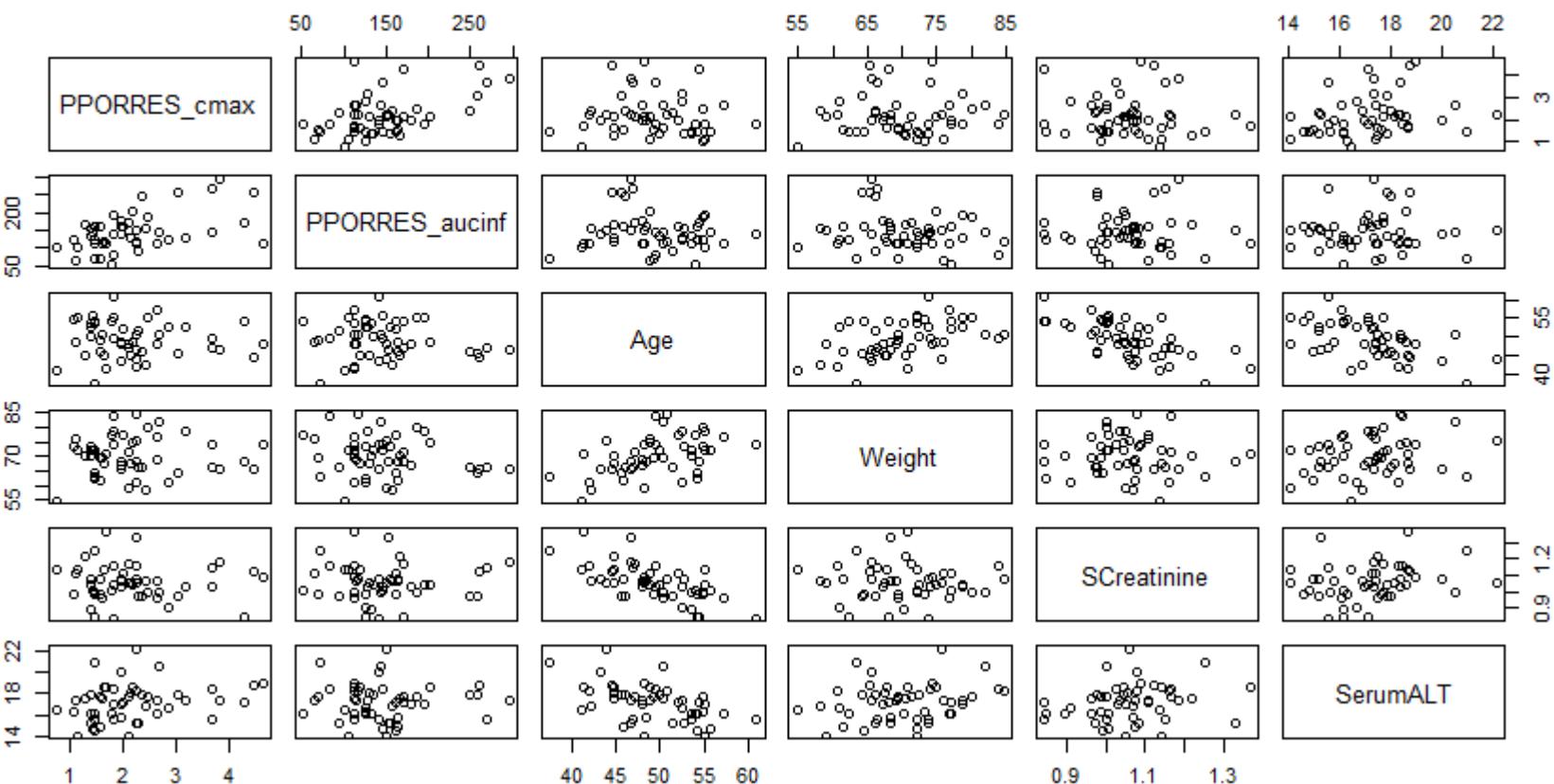
- First, plot data generally using a box scatter plot
- Comparing PK exposure by demographic data

```
### scatter plot matrix for broad overview ###
library(lattice)
# Basic Scatterplot Matrix for exposure vs continuous variables only---
pairs(~PPORRES_cmax+PPORRES_aucinf+Age+Weight+
      SCreatinine+SerumALT,data=alldata,
      main="Scatterplot Matrix for exposure vs continuous variables only")

pairs(~PPORRES_cmax+PPORRES_aucinf+Race+Gender ,data=alldata,
      main="Scatterplot Matrix for exposure vs continuous variables only")

pairs(~PPORRES_cl+Age+Weight+SCreatinine+SerumALT,data=alldata,
      main="Scatterplot Matrix for clearance vs continuous variables only")
```

Scatterplot Matrix for exposure vs continuous variables only



PK Correlatives

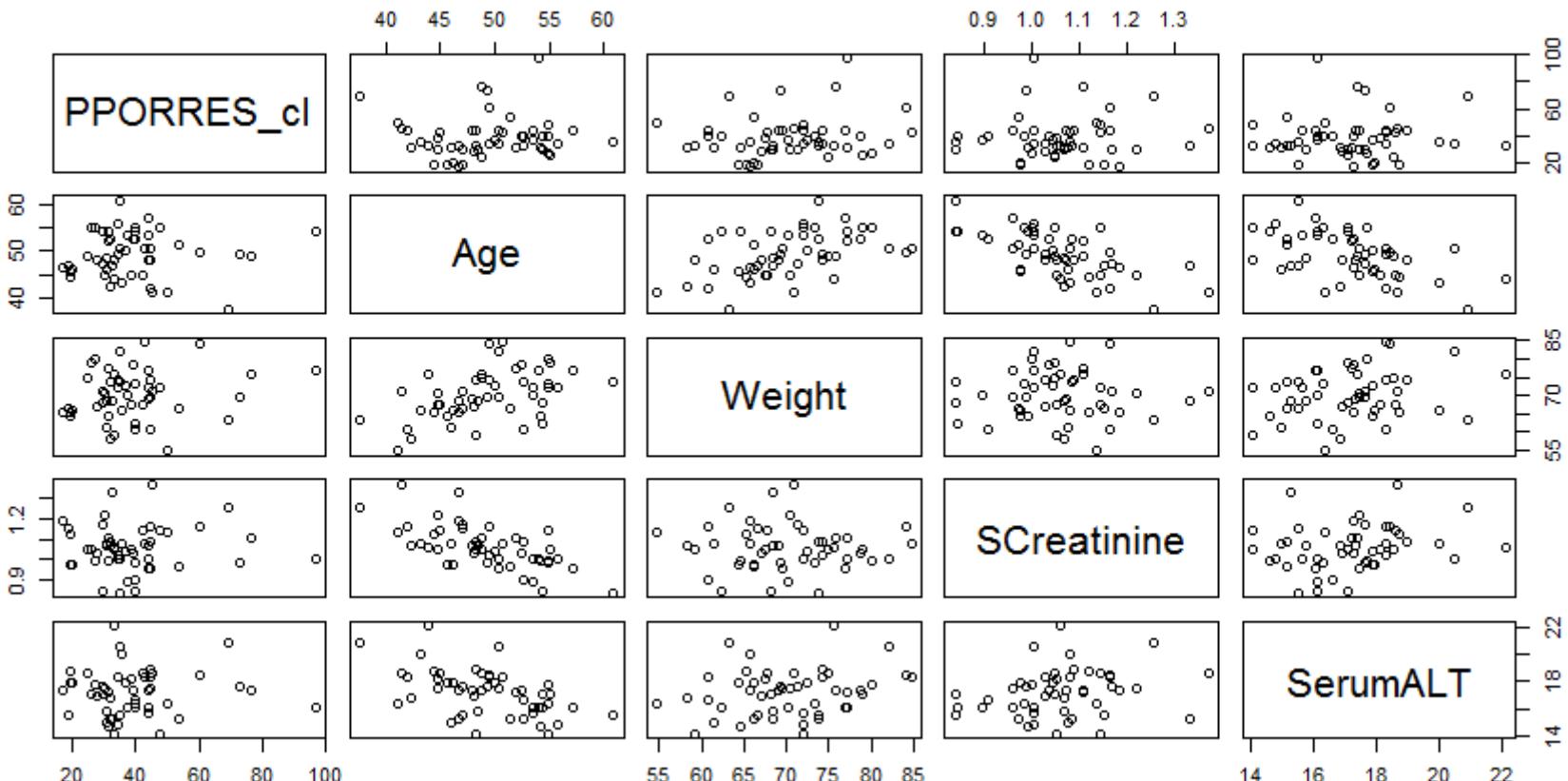
- Comparing Clearance by demographic data

```
### scatter plot matrix for broad overview ###
library(lattice)
# Basic Scatterplot Matrix for exposure vs continuous variables only---
pairs(~PPORRES_cmax+PPORRES_aucinf+Age+Weight+
      SCreatinine+SerumALT,data=alldata,
      main="Scatterplot Matrix for exposure vs continuous variables only")

pairs(~PPORRES_cmax+PPORRES_aucinf+Race+Gender,data=alldata,
      main="Scatterplot Matrix for exposure vs continuous variables only")

pairs(~PPORRES_cl+Age+Weight+SCreatinine+SerumALT,data=alldata,
      main="Scatterplot Matrix for clearance vs continuous variables only")
```

Scatterplot Matrix for clearance vs continuous variables only

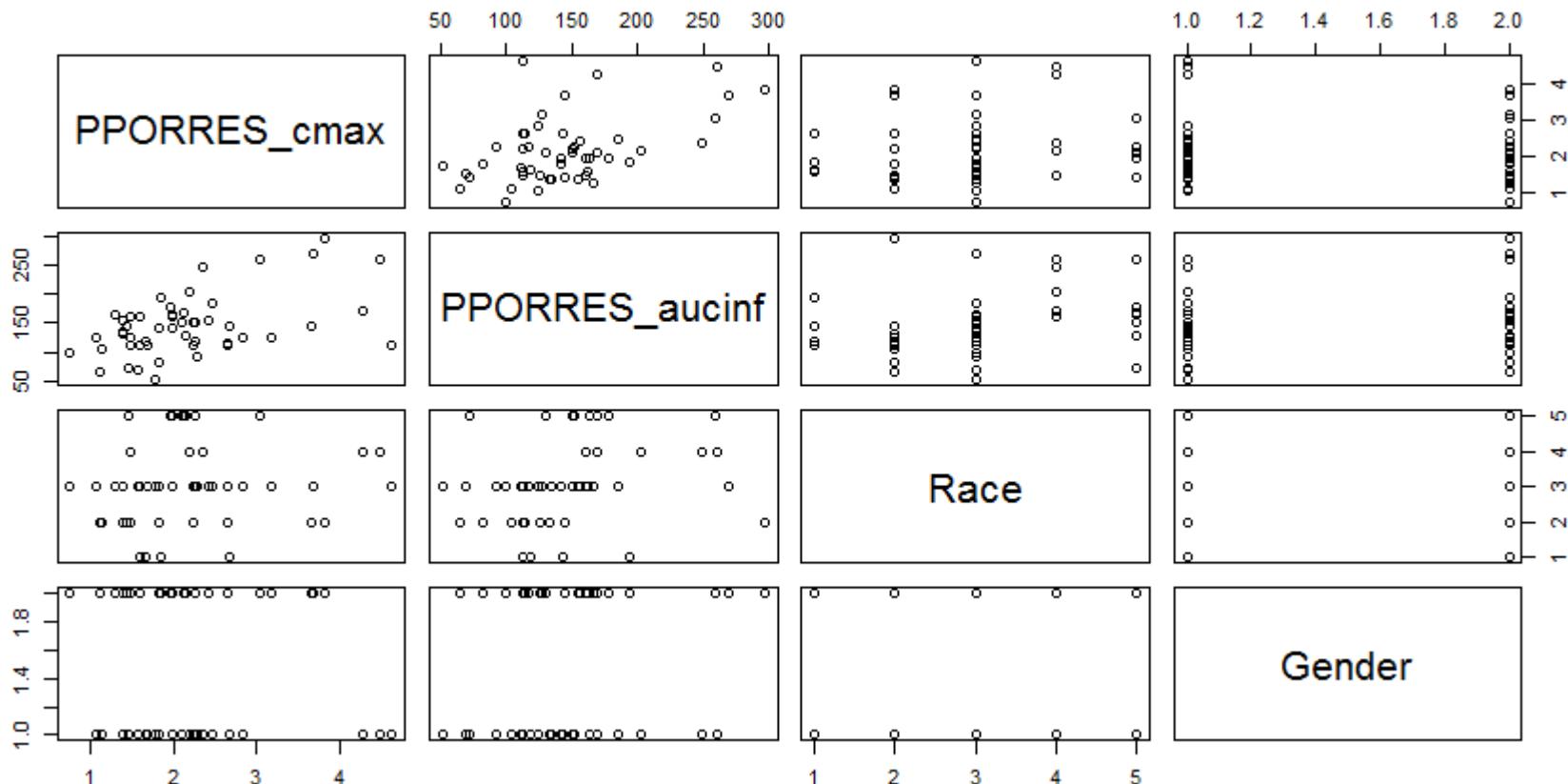


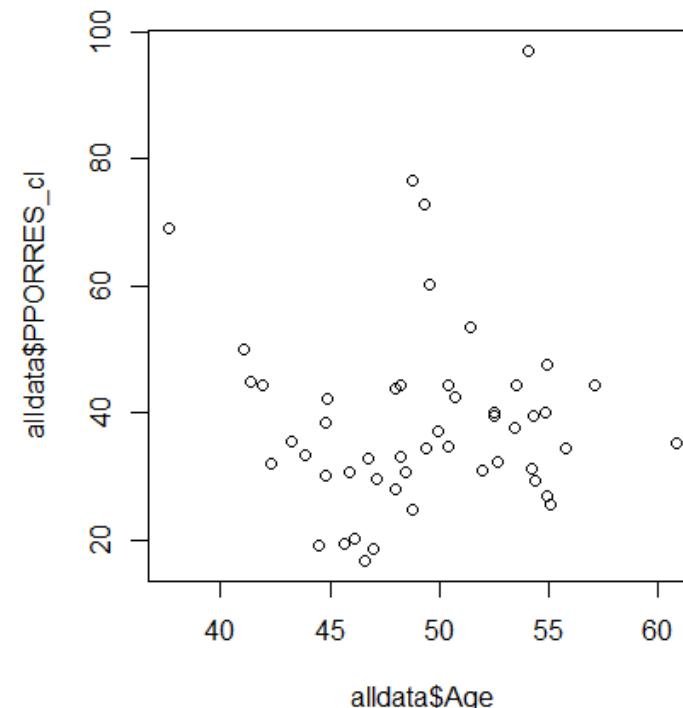
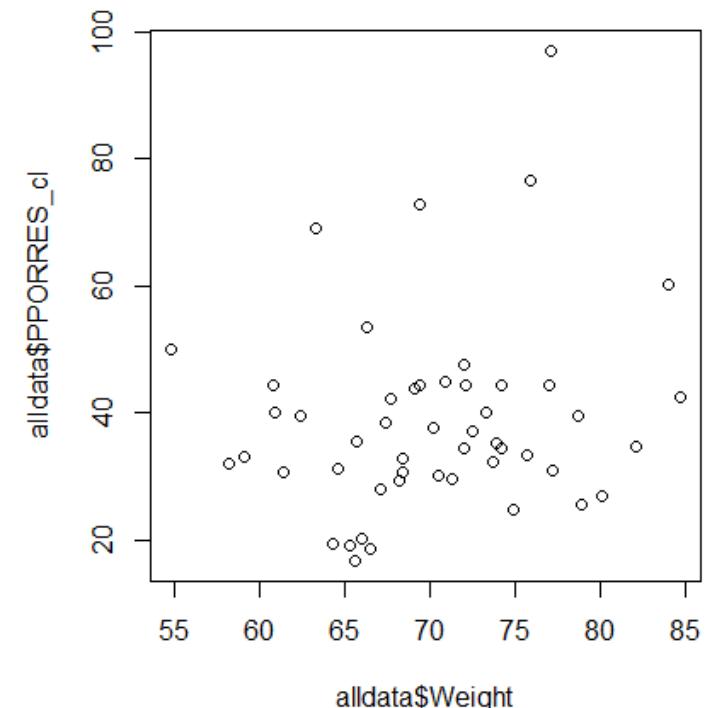
PK Correlatives

- Comparing Clearance by demographic data

```
### scatter plot matrix for broad overview ###
library(lattice)
# Basic Scatterplot Matrix for exposure vs continuous variables only---
pairs(~PPORRES_cmax+PPORRES_aucinf+Age+weight+
      SCreatinine+SerumALT,data=alldata,
      main="Scatterplot Matrix for exposure vs continuous variables only")
pairs(~PPORRES_cmax+PPORRES_aucinf+Race+Gender ,data=alldata,
      main="Scatterplot Matrix for exposure vs continuous variables only")
pairs(~PPORRES_c1+Age+Weight+SCreatinine+SerumALT,data=alldata,
      main="Scatterplot Matrix for clearance vs continuous variables only")
```

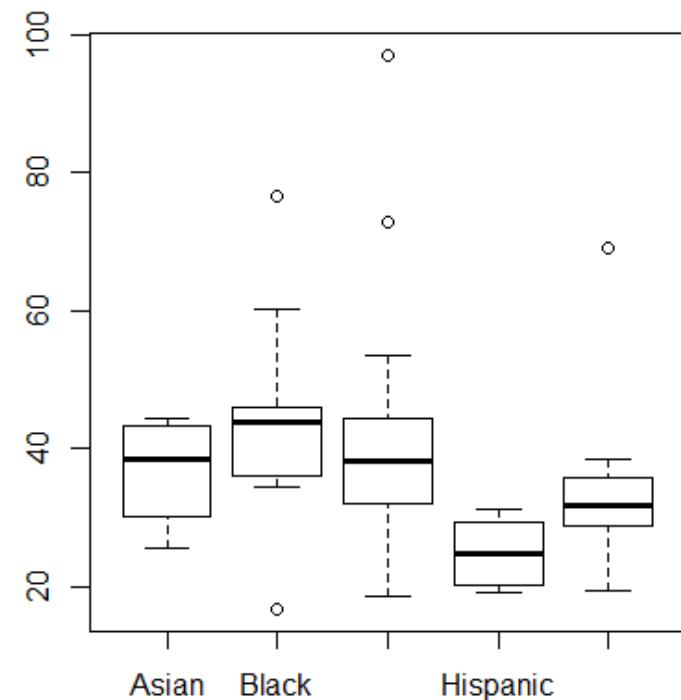
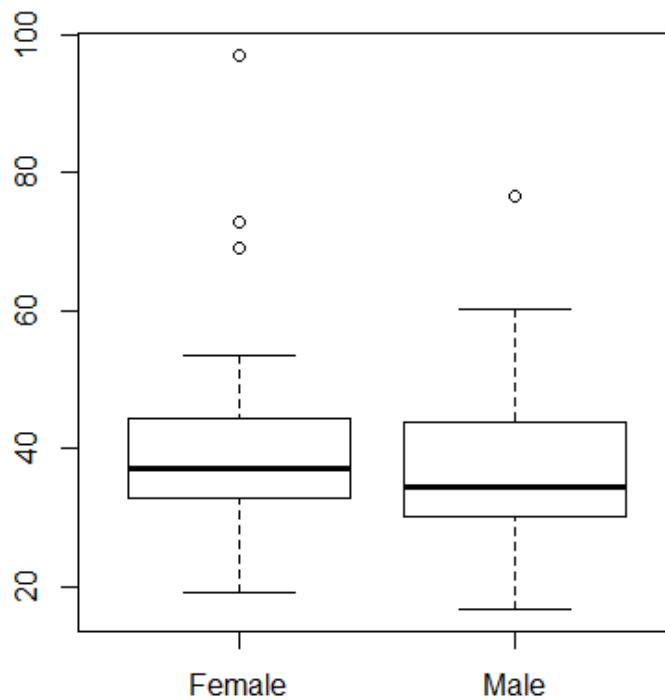
Scatterplot Matrix for exposure vs ordinal variables only

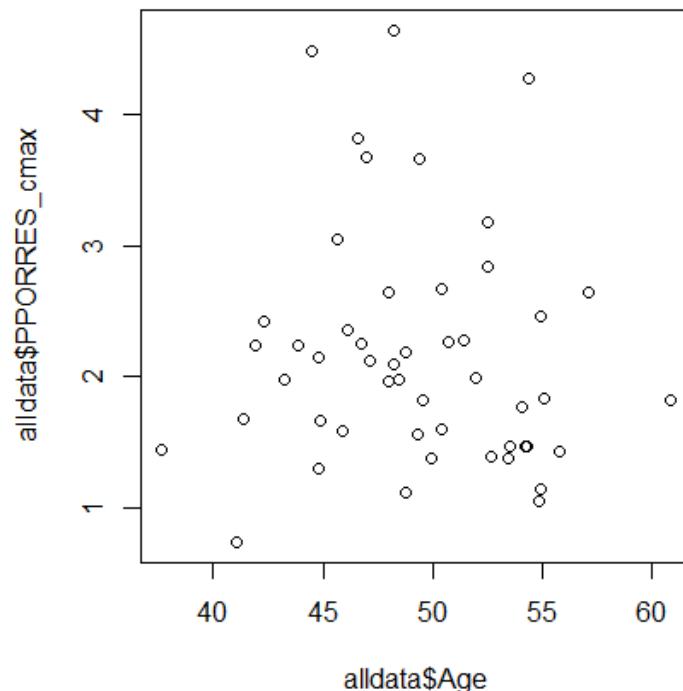
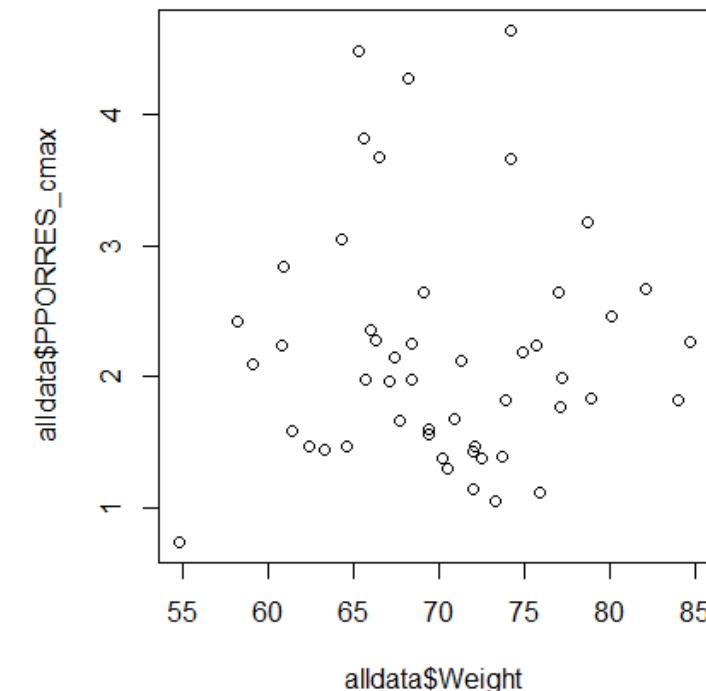




```
par(mfrow=c(1,2))
plot(alldata$Weight, alldata$PPORRES_cl)
plot(alldata$Age, alldata$PPORRES_cl)
```

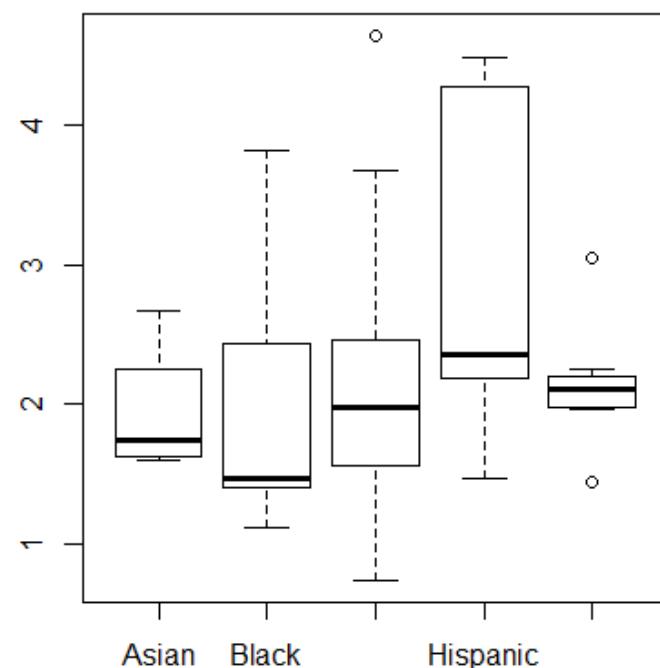
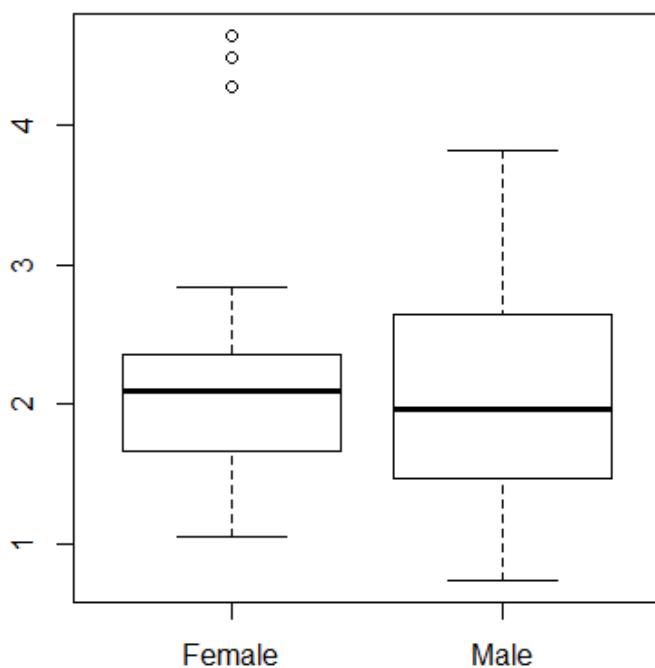
```
par(mfrow=c(1,2))
plot(alldata$Gender, alldata$PPORRES_cl)
plot(alldata$Race, alldata$PPORRES_cl)
```

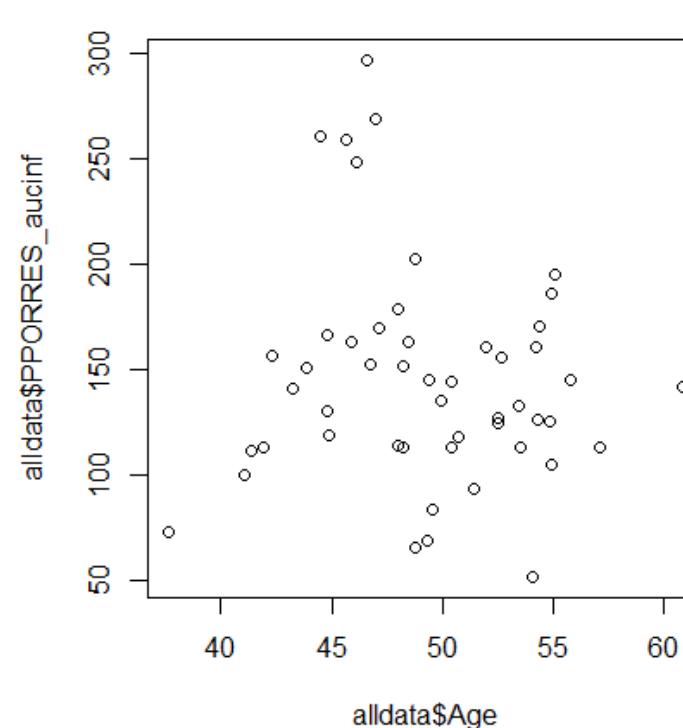
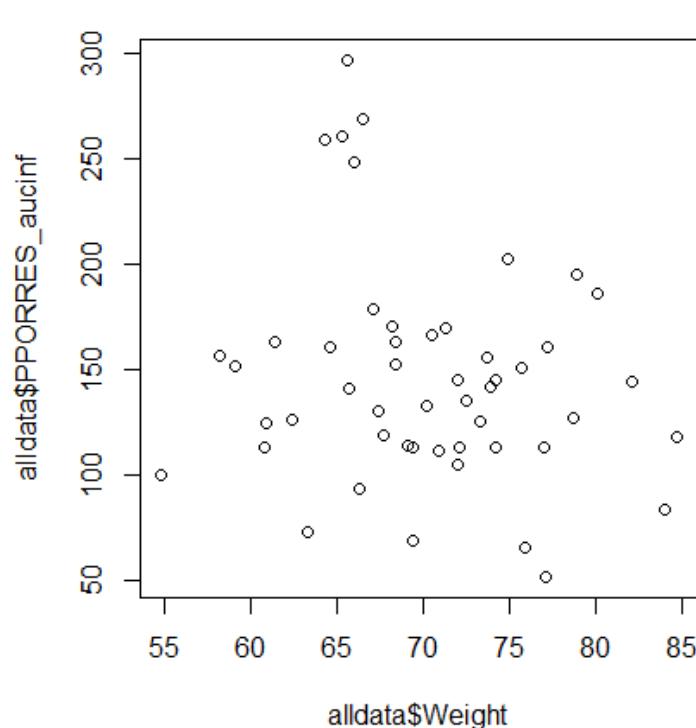




```
par(mfrow=c(1,2))
plot(alldata$Weight, alldata$PPORRES_cmax)
plot(alldata$Age, alldata$PPORRES_cmax)
```

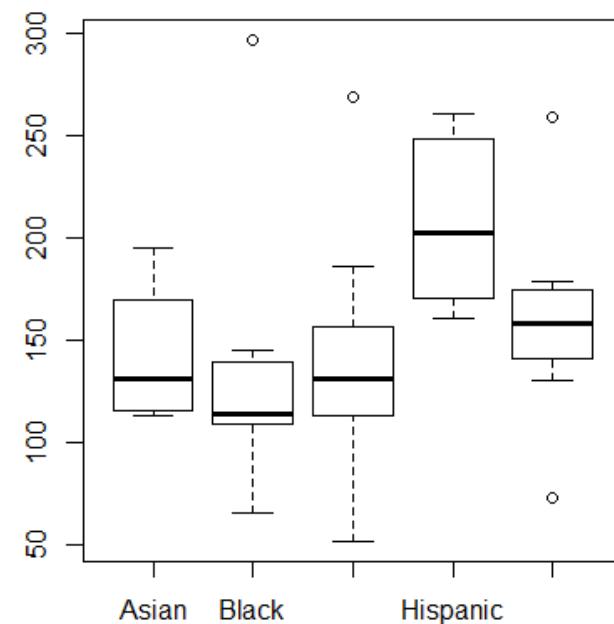
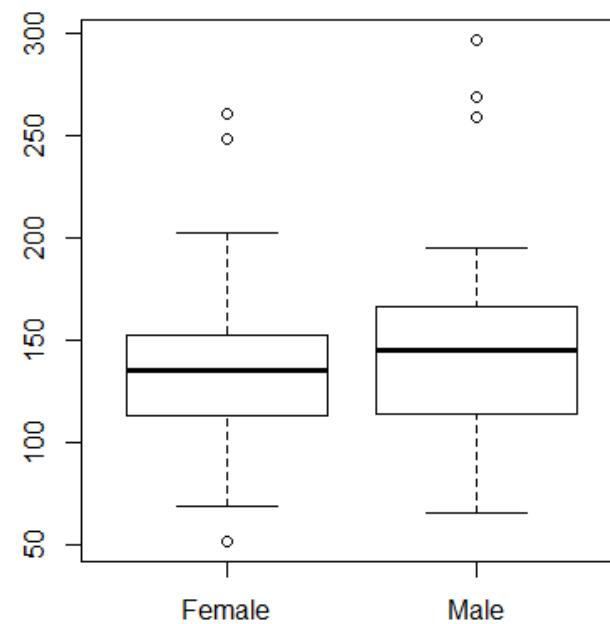
```
par(mfrow=c(1,2))
plot(alldata$Gender, alldata$PPORRES_cmax)
plot(alldata$Race, alldata$PPORRES_cmax)
```





```
par(mfrow=c(1,2))
plot(alldata$Weight, alldata$PPORRES_aucinf)
plot(alldata$Age, alldata$PPORRES_aucinf)
```

```
par(mfrow=c(1,2))
plot(alldata$Gender, alldata$PPORRES_aucinf)
plot(alldata$Race, alldata$PPORRES_aucinf)
```



Alternatives

- If prefer to analyze correlatives in different software (Excel®, GraphPad Prism®, etc), can **export** the “alldata” dataframe outside of R in a .csv file format
 - Uses the “write.csv” function
 - First, specify the R dataframe, then name the file
 - File will appear in the Working Directory

```
write.csv(alldata, file="FinalPKNCA.csv")
```

PK Summary

- Performing exposure/response analyses requires calculating the drug exposure
 - Cmax, AUC are most common metrics for drug exposure
- Can calculate AUC using “PKNCA” package in R for free
- Using noncompartmental methods, can also calculate secondary PK parameters (based off the AUC), such as clearance and volume of distribution
- Can use these parameters to correlate with your pharmacodynamic (PD) response outcomes

PM Break

Clinical Trials

Clinical Trials

- Ultimate goal is to demonstrate to the FDA that your drug is safe and effective
- Give 1000s of patients the drug
 - Monitor safety (toxicity) and efficacy compared to placebo or SOC
- Approval based on *totality of evidence*
 - FDA must be convinced beyond a doubt that drug (at given dose/schedule) is safe and effective
 - Costly (drug sponsor pays for drug given to patients)

Clinical Trials

- Before drug companies spend millions on several large trials where each patient given same dose, must optimize that dose
- Try different dosing routes, dose amounts, dose frequencies in a few small trials to grasp optimal dosing based on PK
 - Can be done in patients, healthy volunteers, or both
- Then scale up size of trial in patients, trying either a single dose level or 2-3 dose levels
 - Look for evidence of efficacy

Types of Clinical Trials

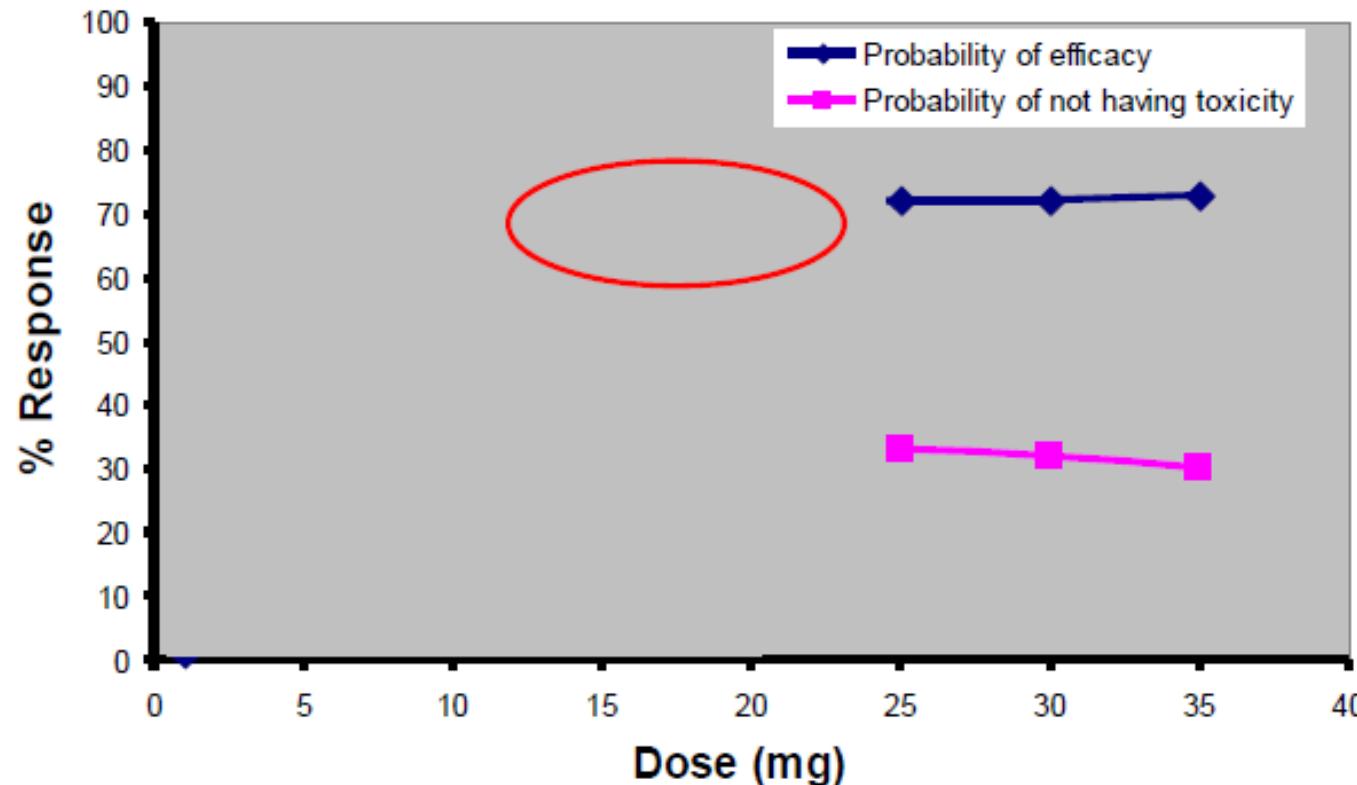
- Phase 0: first-in-human
 - Confirm drug hitting target
- Phase I: dose escalation
 - PK, MTD, small enrollment (typically $n < 30$)
- Phase II: efficacy, toxicity assessment
 - Additional PK, larger enrollment (typically $30 < n < 100$)
- Phase III: confirmatory
 - Large enrollment ($n > 100s-1000s$)
- Phase IV: post-marketing

Preclinical Drug Development

- Pharmacokinetic and efficacy studies performed in rodents, possibly larger animals
 - Both healthy and disease bearing (if possible, e.g. tumor xenografts)
 - Weight-normalized dosing
- Assays developed and validated
 - PK assays
 - LC-MS/MS for small molecules
 - ELISA for biologics
 - Pharmacodynamic (PD) assays
 - Depending on what the target and mechanism is, can design assay to assess drug effect
 - E.g. inhibition of target mechanism (measure decrease in “products”)
- By end of this stage, scientists should have good idea of what drug exposure is needed to elicit a desired effect
 - Dose (mg/kg) should be known that is needed to achieve effective exposure (PK studies)

Selection of a First-in-Humans Dose

- Dose means everything
 - “Only dose determines the poison” – Paracelsus (1493-1541)
 - Even water is lethal if given enough (dose-dependent)
 - Finding optimal dose for efficacy and safety is essential for successful drug development



Selection of a First-in-Humans Dose

- FDA has guidance for deriving the maximum recommended starting dose (MRSD)
 - Based on preclinical doses given, observed toxicities, and an algorithm for MRSD
 - The most common approach
- There is an alternative approach that uses animal PK/PD data modeling to select a dose
 - Relatively rare, as most animal studies have insufficient data for this type of derivation

Selection of a First-in-Humans Dose

- Using most common method involving animal dose/toxicity relationships
 - Determine the dose where no observed adverse effect levels are seen (NOAEL) in all animal species tested
 - For small molecules only
 - For biologics, use minimum anticipated biological effect level (MABEL)
 - Once NOAEL dose is known, convert that to a human-equivalent dose (HED) based on allometric scaling
 - A species specific conversion factor (CF) to convert that animal's NOAEL to the HED
 - CF based on normalizing body surface area (BSA) for that animal species
 - Converts animal mg/kg dose to a human mg/kg dose

Table 3: Conversion of Animal Doses to Human Equivalent Doses Based on Body Surface Area

Species	Reference Body Weight (kg)	Working Weight Range ^a (kg)	Body Surface Area (m ²)	To Convert Dose in mg/kg to Dose in mg/m ² Multiply by k _m	To Convert Animal Dose in mg/kg to HED ^b in mg/kg. Either	
					Divide Animal Dose By	Multiply Animal Dose By
Human	60	---	1.62	37	---	---
Child ^c	20	---	0.80	25	---	---
Mouse	0.020	0.011-0.034	0.007	3	12.3	0.081
Hamster	0.080	0.047-0.157	0.016	5	7.4	0.135
Rat	0.150	0.080-0.270	0.025	6	6.2	0.162
Ferret	0.300	0.160-0.540	0.043	7	5.3	0.189
Guinea pig	0.400	0.208-0.700	0.05	8	4.6	0.216
Rabbit	1.8	0.9-3.0	0.15	12	3.1	0.324
Dog	10	5-17	0.50	20	1.8	0.541
Primates:						
Monkeys ^d	3	1.4-4.9	0.25	12	3.1	0.324
Marmoset	0.350	0.140-0.720	0.06	6	6.2	0.162
Squirrel monkey	0.600	0.290-0.970	0.09	7	5.3	0.189
Baboon	12	7-23	0.60	20	1.8	0.541
Micro-pig	20	10-33	0.74	27	1.4	0.730
Mini-pig	40	25-64	1.14	35	1.1	0.946

^a For animal weights within the specified ranges, the HED for a 60 kg human calculated using the standard k_m value will not vary more than ± 20 percent from the HED calculated using a k_m value based on the exact animal weight.

^b Assumes 60 kg human. For species not listed or for weights outside the standard ranges, human equivalent dose can be calculated from the formula:

$$\text{HED} = \text{animal dose in mg/kg} \times (\text{animal weight in kg}/\text{human weight in kg})^{0.33}$$

^c The k_m value is provided for reference only since healthy children will rarely be volunteers for phase 1 trials.

^d For example, cynomolgus, rhesus, and stump-tail.

Selection of a First-in-Humans Dose

- Starting dose in humans = HED/safety factor
- Most common safety factor value is 10
 - Ensures a SAFE starting dose
 - Whole goal of a first-in-humans dose is to enter clinical studies with no toxicity

Clinical Trial Design – Phase 0

- A first-in-humans clinical trial
- Conducted prior to a dose-escalation (Phase 1) trial
- AKA Pilot studies
- Can commence before all the preclinical toxicity data has been gathered
- Helps identify drugs with poor PK or human *in vivo* activity earlier
 - Makes for more efficient clinical drug development

Clinical Trial Design – Phase 0

Typical Objectives

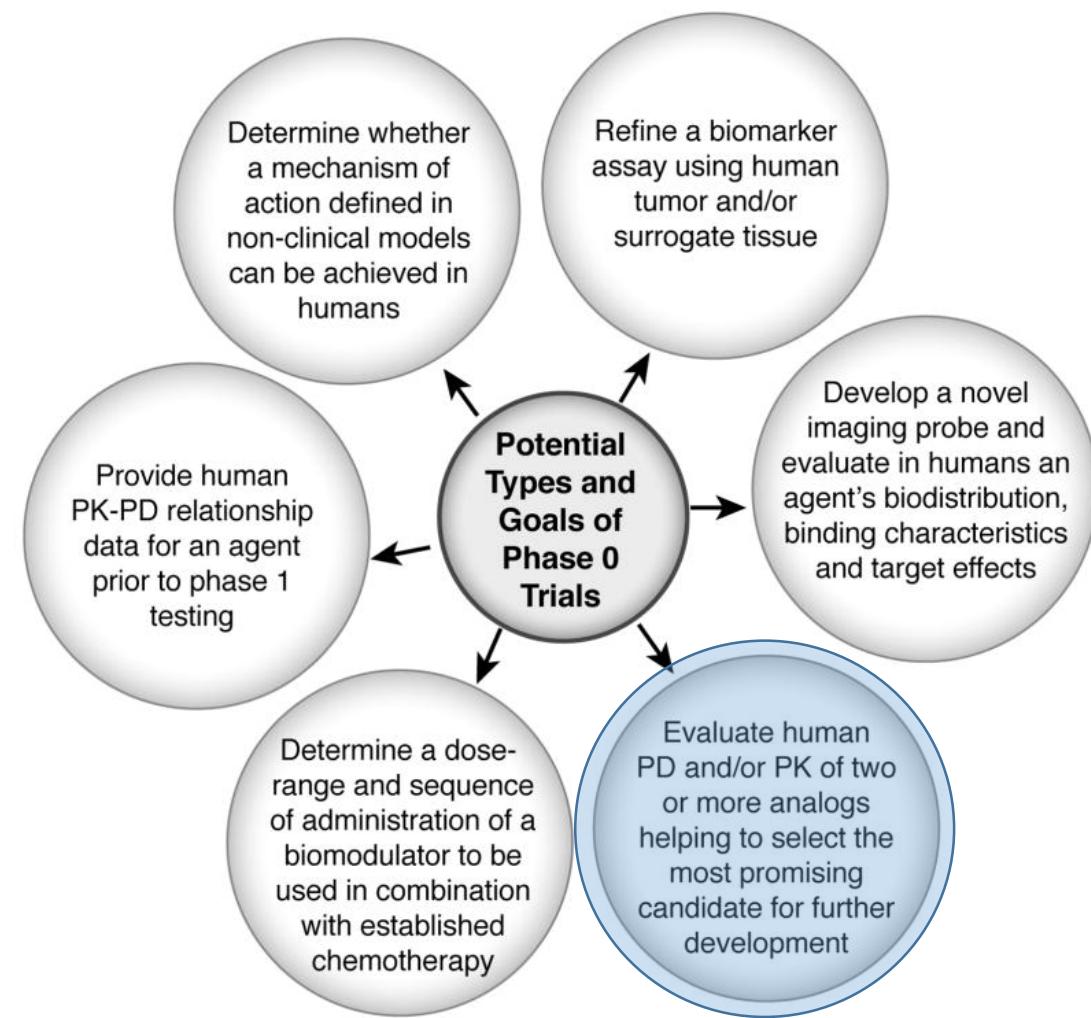
1. Validated preclinical PD assays
 2. Ensure drug is hitting target/inducing effect
-
- Give low doses for short period of time
 - Exposure levels should be predictable based on preclinical PK studies
 - No chance for therapeutic benefit (dose and duration of treatment too short)
 - Qualifies for FDA Exploratory IND Guidance
 - Initiating a Phase 0 requires less preclinical tox data than a Phase 1 would

Phase 0

- Ideal drug candidates for Phase 0 testing:
 1. Successful clinical development depends heavily on a PD endpoint
 - i.e. a drug who exerts effect on a known target involved in a known mechanism, and ability to measure target activity/mechanism activity is crucial to understanding drug exposure/response relationship
 2. Drug's target is “credentialed”
 - Preclinical studies showed modulation of this target is correlated/associated with efficacy (e.g. tumor shrinkage)
 3. A wide therapeutic window
 - Exposures can vary quite a bit while being therapeutic but not toxic
 4. Drug's target can be modulated at low (non-toxic) doses given for short duration
 5. Target modulation can be determined with a small patient set (n= 10-15)

Phase 0 Trial Design

- Preclinical PK studies typically performed in small rodents initially
 - Half-life
 - Optimal route, dose (mg/kg)
 - Bioavailability
- Sometimes scaled up to larger rodents or larger mammals
 - Pigs, dogs, monkeys
- Phase 0 can be used to assess PK in humans
 - Identify possible allometric (body size-based) scaling
 - Useful for pediatrics, neonates, etc
 - Optimize formulation, route, dose frequency

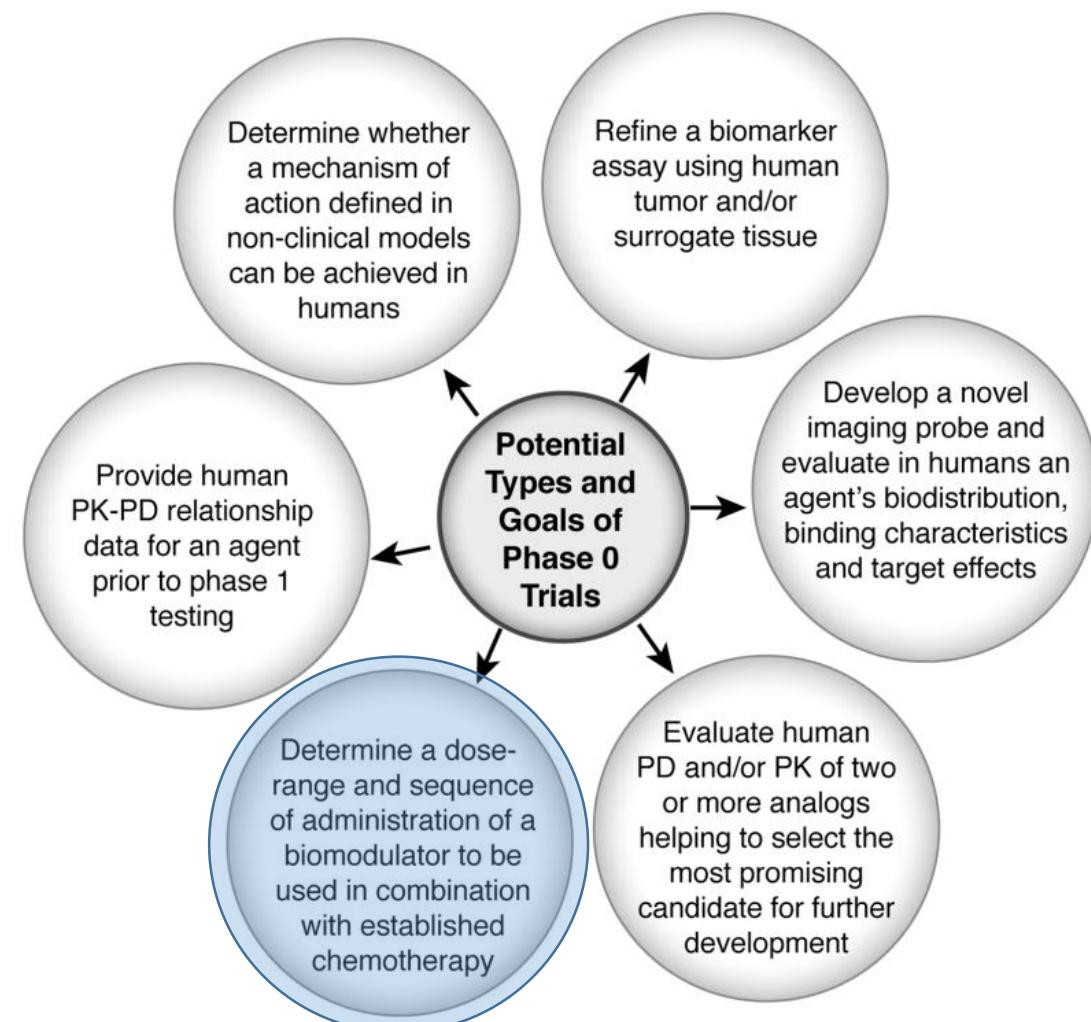


Murgo et al. *Clin Cancer Res.* 14(12):3675-82 (2008)

$$Y = a W^b$$

Phase 0 Trial Design

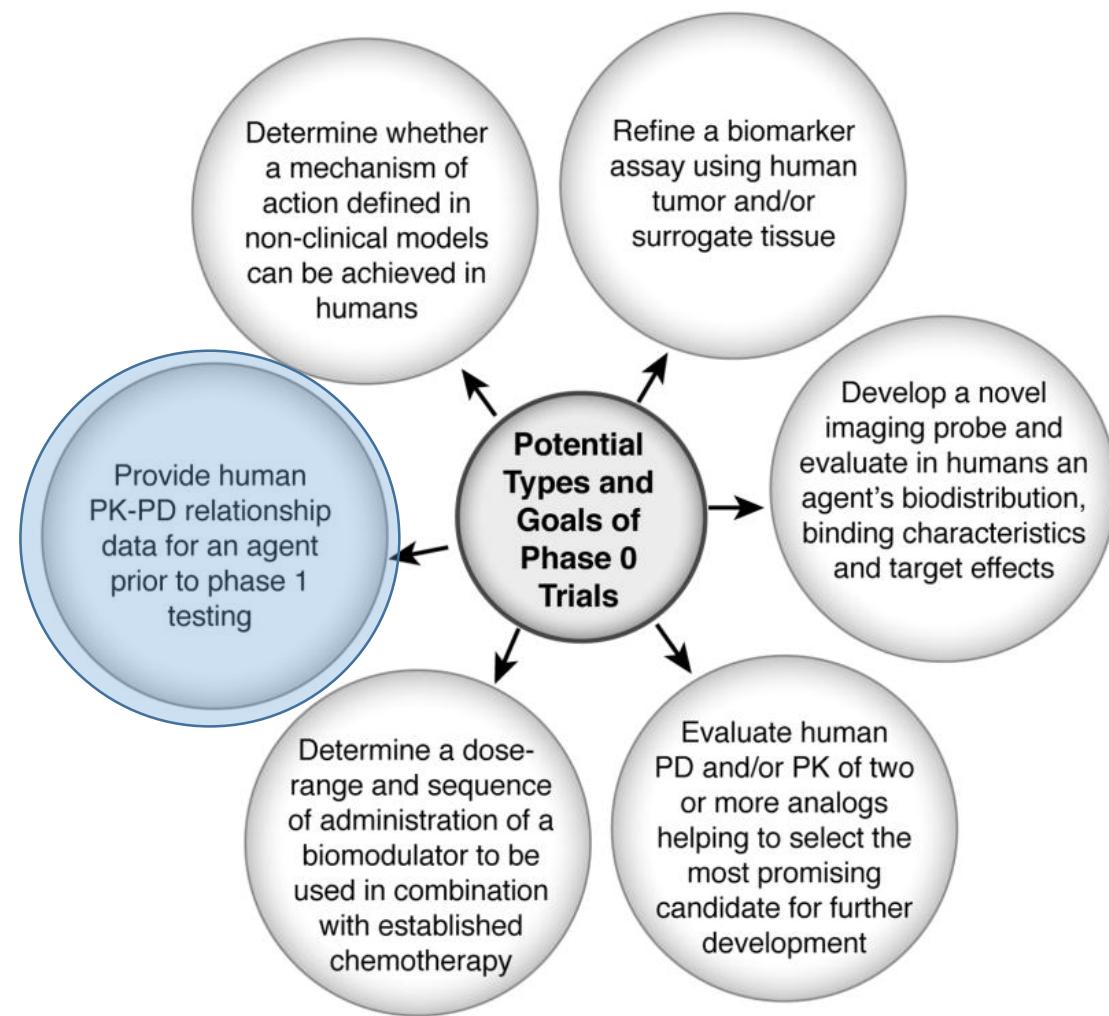
- Only if drug is to be given with another agent or biomodulator
- Biomodulator used to alter drug's ADME or to enhance efficacy
 - Ex. 5-azacytidine is effective, but quickly degrade by cytidine deaminase enzyme. THU blocks this enzyme. So 5-aza given with THU
- Phase 0 can be used to optimize the sequence of administration, dose, route, frequency



Murgo et al. Clin Cancer Res. 14(12):3675-82 (2008)

Phase 0 Trial Design

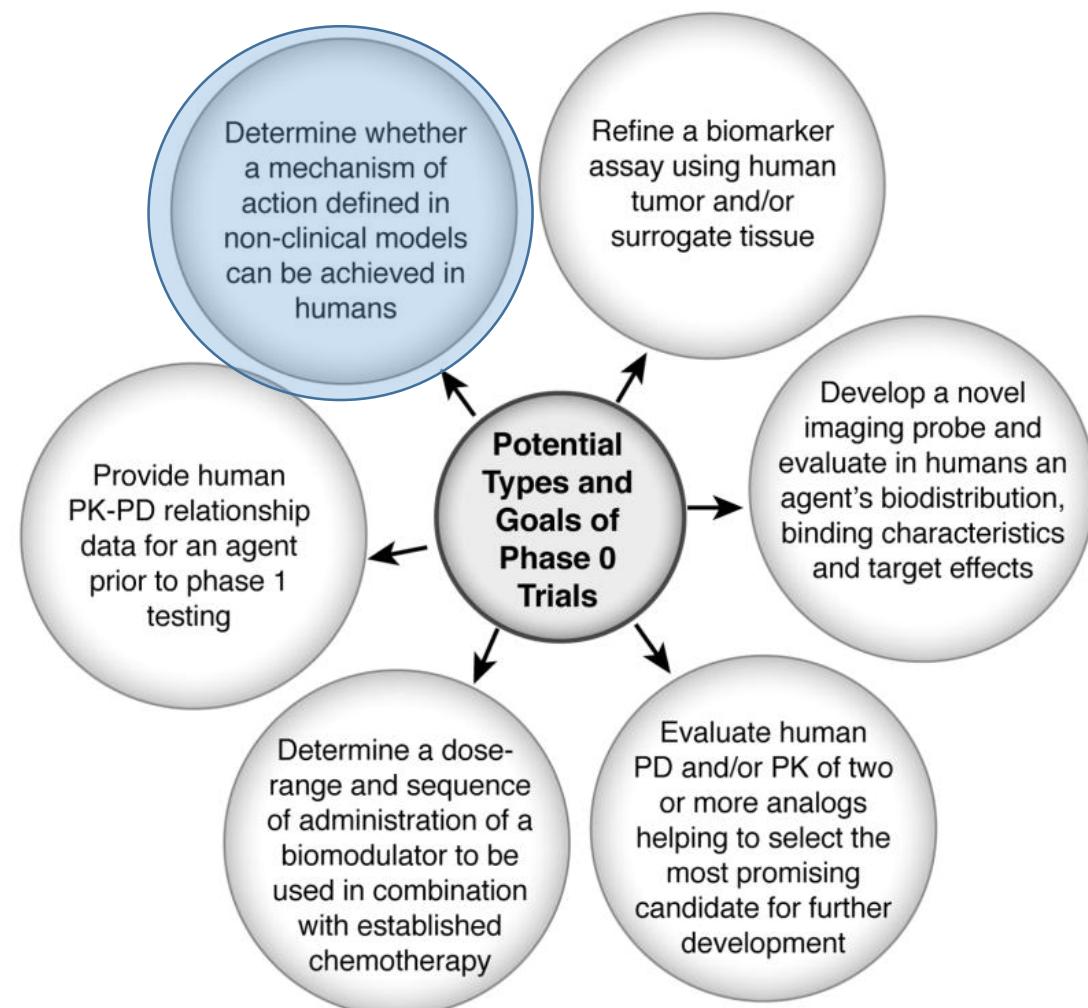
- Preclinical studies often ensure drug is effective at a maximum tolerated dose. PK studies normally completed as well
- Rarely do preclinical studies model PK/PD relationships
- Phase 0 can be used to develop these PK/PD model to predict what dose might induce what effect and to what extent.
 - Can help reduce the number of phase 1 trials
 - Can help narrow dose escalation range
 - Eliminate sub-therapeutic doses



Murgo et al. *Clin Cancer Res.* 14(12):3675-82 (2008)

Phase 0 Trial Design

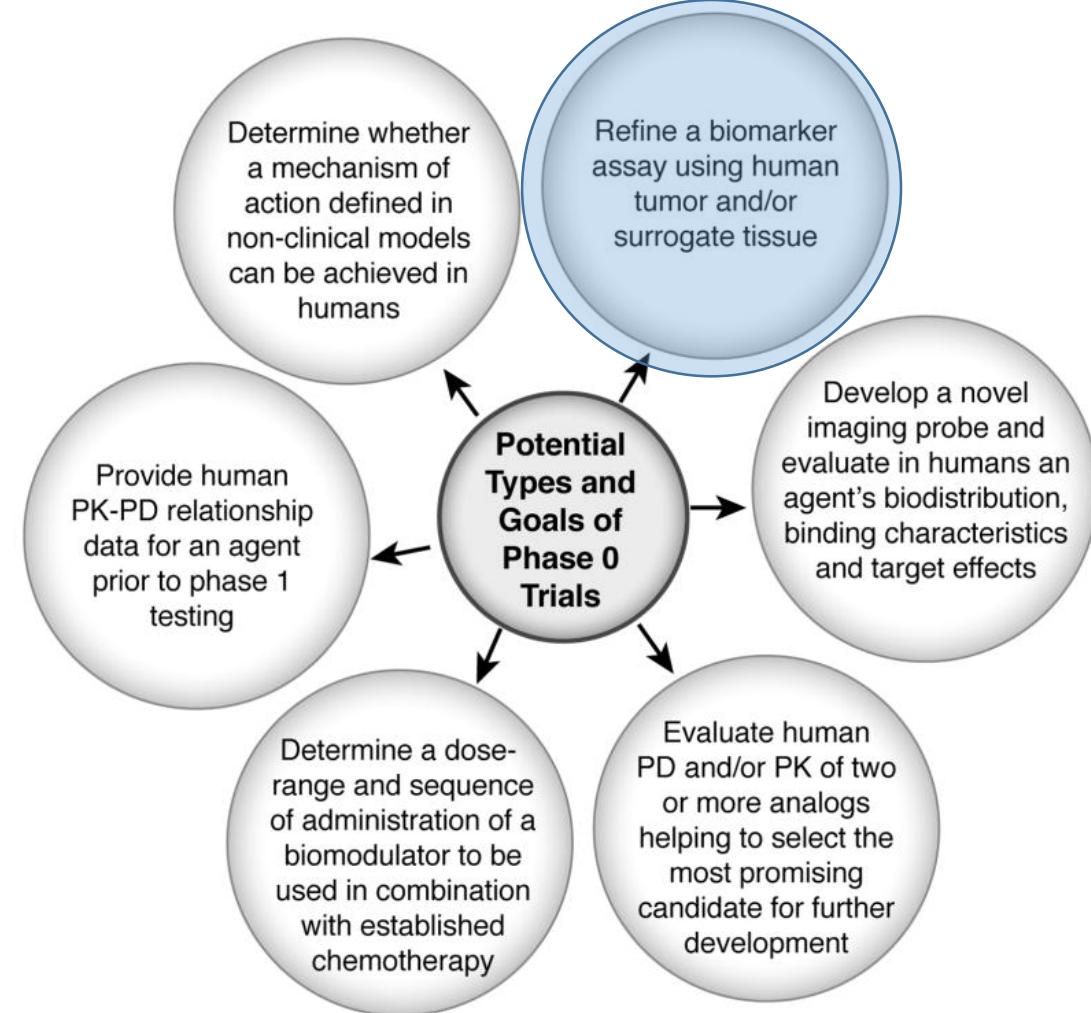
- Preclinical studies identified the drug's target and target's role in larger mechanistic pathway
- Phase 0 can be used to confirm this mechanism occurs in humans and that the drug can modulate similar to preclinical models



Murgo et al. *Clin Cancer Res.* 14(12):3675-82 (2008)

Phase 0 Trial Design

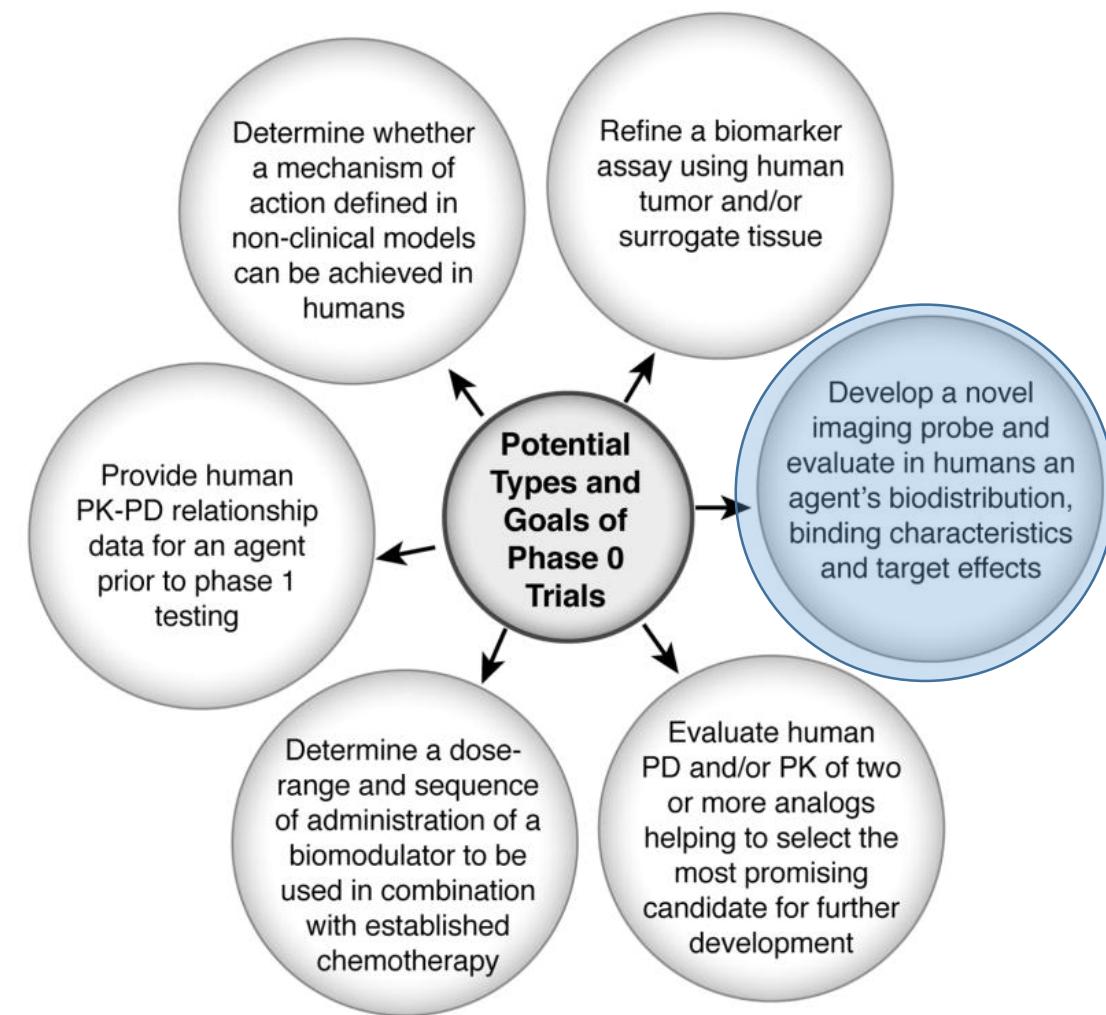
- Preclinical studies identified a biomarker that can measure activity of drug target
 - Directly or indirectly
 - An easy to access biosample that can be processed to measure a specific biomarker
- Assay was designed and/or validated to measure this biomarker
- Phase 0 can be used to confirm and/or refine the biomarker assay
 - Using human tissue or some surrogate of



Murgo et al. Clin Cancer Res. 14(12):3675-82 (2008)

Phase 0 Trial Design

- Preclinical studies unlikely to progress to a point that involves imaging probes
 - Other than radioactively-tagged drug to do PK-based mass-balance study
- Phase 0 can be used to confirm mass balance (aka biodistribution) in humans
- Phase 0 also used to apply imaging probes to “see” drug’s effect in action
 - Changes in imaging status before vs after drug



Murgo et al. *Clin Cancer Res.* 14(12):3675-82 (2008)

Clinical Trial Design – Phase I

- Design based on objective
- Objective:
 1. Determine the maximum tolerated dose of single agent
 2. Determine the optimal route of a single agent
 3. Determine the optimal frequency of a single agent
 4. Determine the sequence and/or dose of a new combination

Clinical Trial Design – Phase I

- Increase dose incrementally with pre-determined dose levels
 - Based on prior clinical data, or if a first-in-humans, based on a model-predicted dose from pre-clinical animal data
 - Once first dose level determined, higher doses are *typically* determined by a Fibonacci or modified Fibonacci sequence:
 1. Initial dose
 2. 2x initial dose
 3. 1.67x previous dose
 4. 1.50x previous dose
 5. 1.33x previous dose
 6. 1.33x previous dose
 - Usually have n=3 subjects/patients on each dose level

Phase 1

- Identify the maximum tolerated dose (MTD)
 - Based on number of patients who experience drug-related dose-limiting toxicities (DLTs)
 - DLT defined as:
 - Any grade 3 adverse event (AE)
 - Excludes hypertension that can be controlled with an anti-hypertensive
 - Any grade 2 AE that persists for 14 days
 - Grade 4+ neutropenia that lasts for at least 5 days
 - Other specifics are trial- and drug-dependent
- If 0/3 subjects have no DLTs, then progress to next dose level
- If 1/3 subjects have DLTs, then treat another 3 subjects at same dose level
 - If ≤ 1 of 6 subjects have DLTs, then can progress to next dose level
- If 2/3 subjects have DLTs, then *previous* dose level considered to be MTD
 - MTD usually the recommended phase 2 dose (RP2D)

Clinical Trial Design – Phase 2

- Confirm PK in a larger cohort using patients that actually need the drug
 - As opposed to healthy volunteers in Phase 0-1
 - Disease states may alter drug PK
 - Liver and kidney cancer may have slower drug CL, greater exposure, greater risk of AEs
- Objectives:
 1. Determine whether drug offers therapeutic benefit (e.g. survival improvement)
 2. Measure the PD (biomarkers) to ensure drug doing what it was designed to do on a molecular level (e.g. kinase inhibition)
 3. Provide additional PK and safety (toxicity) data
- Assess efficacy in a particular patient population
 - Using the recommended phase 2 dose (RP2D)

Clinical Trial Design – Phase 2

- Usually a single arm study
 - Give x number of patients the same dose/schedule
 - Monitor efficacy and tox
- Some trials are multiple arms for comparing two things
 - Standard Of Care (SOC) or placebo vs test drug
 - Oral vs Non-oral administration of a drug (bioavailability study)
 - Two different formulations of a drug (bioequivalence study)
 - Single agent vs combination of drugs
 - Optimizing the sequence of administering combo of drugs
- Randomly assign patients to a particular arm of the study (randomized trial)
- Can be blinded to what arm they're on or not

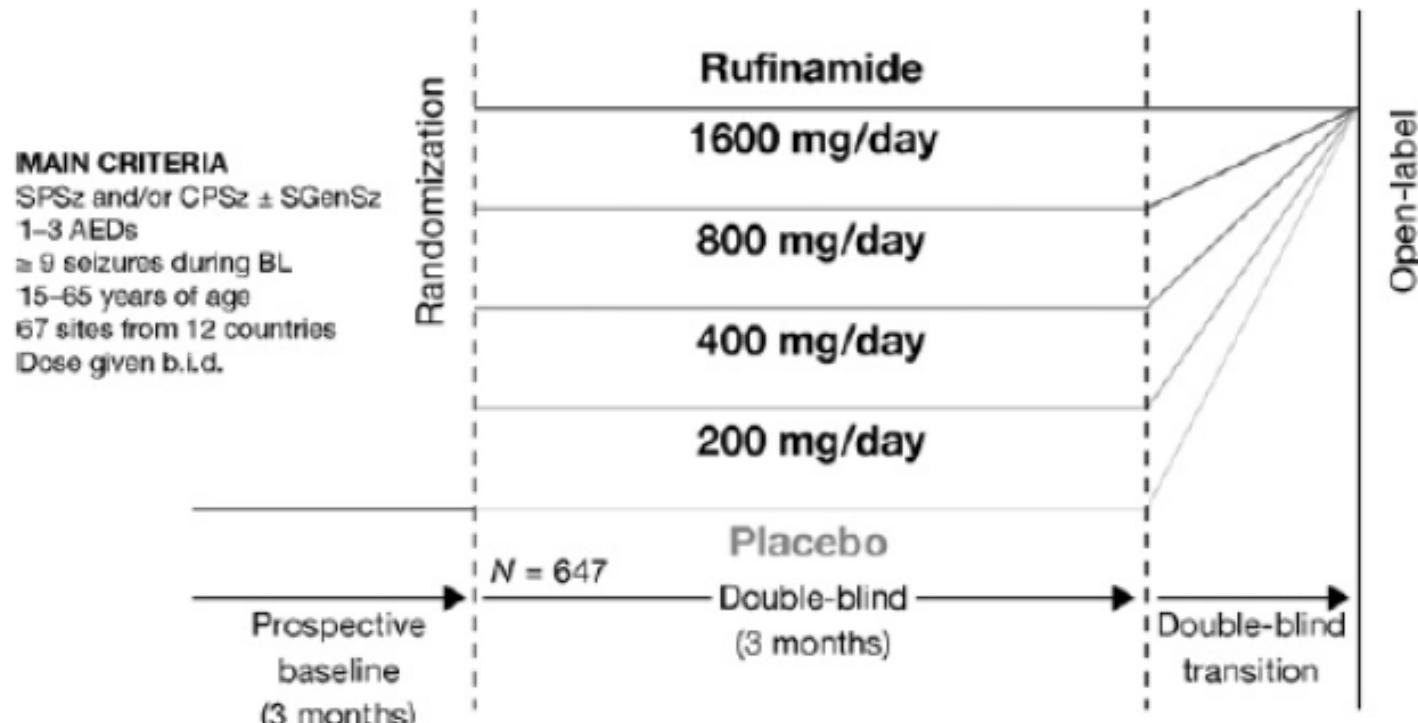
Clinical Trial Design

- Parallel (two arm: randomized to placebo vs treatment; or SOC vs treatment)
 - Superiority
 - Equivalence or Non-Inferiority
 - Dose-Response Relationships
- Crossover (two arm: randomized to an arm, then switch to other arm mid-trial)
 - Bio-equivalence
- Adaptive Design
 - Can modify dosing and/or endpoint criteria mid-trial
 - Maximizes efficiency; minimizes # patients needed to show statistical significance

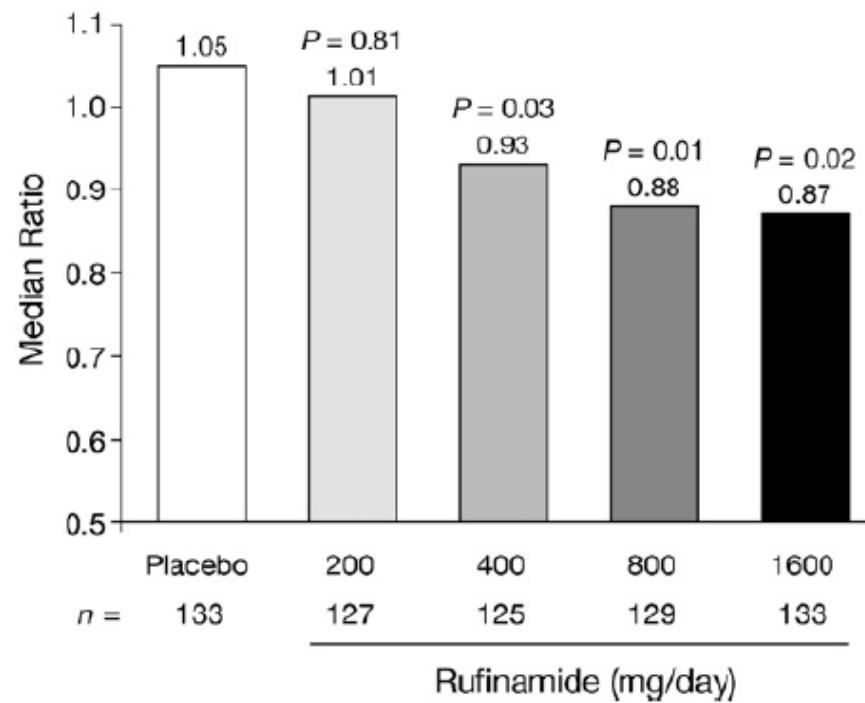
Parallel Design

- Fixed dose trial
 - Patients receive a single dose level (fixed dose) throughout study
 - There may be multiple dose levels in study, but each individual patient would stay on whatever dose they were assigned/randomized to
- Forced titration trial
 - Patients receive gradually higher doses (titration) without regard to response, only tolerability based on adverse events (AE) / toxicity
 - Common Phase I trial design
- Titration to endpoint trial
 - Patients receive gradually higher doses in order to achieve a certain target

Fixed Dose Design - Example

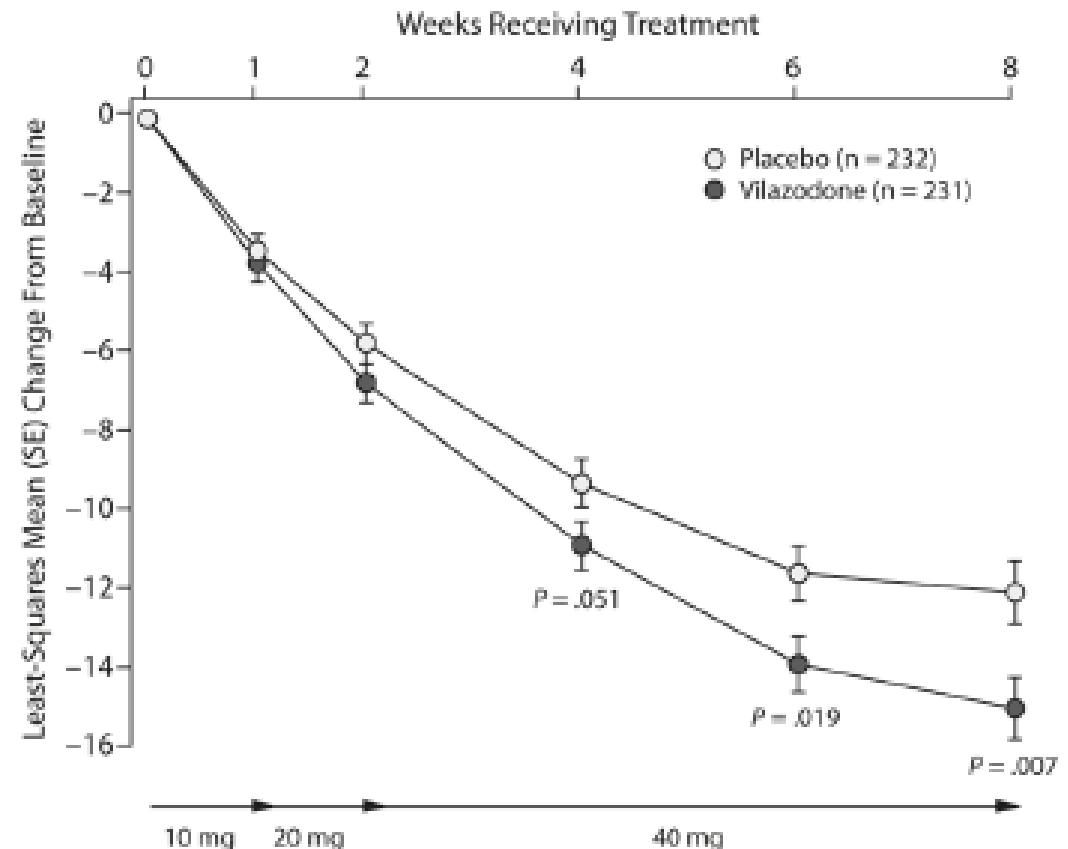


Linear trend for dose response was established; $P = 0.003$



Forced Titration Design - Example

- Week 1: 10 mg QD
- Week 2: 20 mg QD
- Weeks 3-8: 40 mg QD

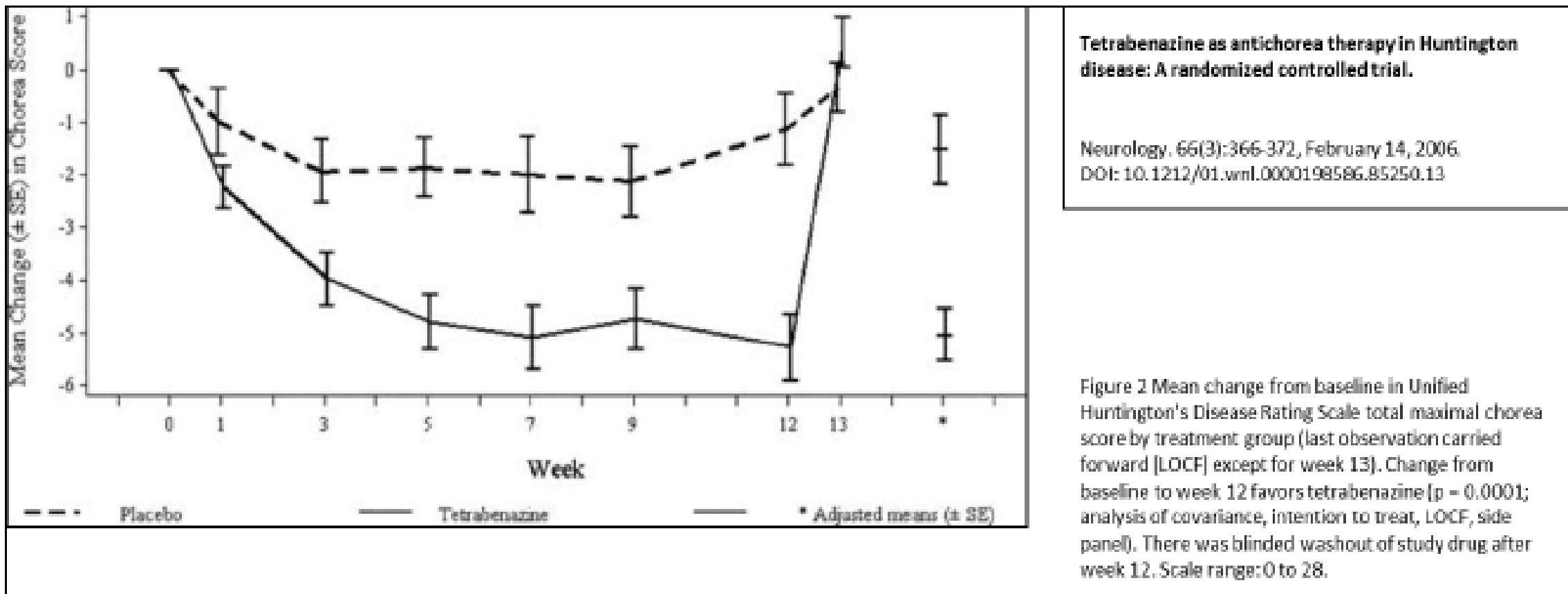


Abbreviations: ITT = intent to treat, MADRS = Montgomery-Asberg Depression Rating Scale, MMRM = mixed-effects model repeated-measures, SE = standard error.

Titration to Endpoint Design

- Individually titrate the dose in each patient based on response measurement
- Example: tetrabenazine
- Formulated in 12.5mg tablets
 - Titrated over first 7 weeks of study
 - Day 1: 1 tablet/day
 - Days 2-7: 2 tablets/day (1 tablet, twice a day)
 - Weeks 2-7: increased 1 tablet/day per week up to 8 tablets/day
 - Until desired effect reached or intolerable AE

Titration to Endpoint Design



Tetrabenazine eventually FDA approved with following dose schedule:

Week 1: 12.5 mg QD

Week 2: 12.5 mg BID (25 mg daily)

Weeks 3+: increase by 12.5 mg/day in weekly intervals

Cross-Over Design

- Two (or more) groups structured into design
- After a period of time, subjects in one group are re-assigned to another group
 - The cross over into another group
- 2 advantages over a parallel design
 1. Each patient serves as their own control (pair-wise comparisons)
 - When same patient experiences both test and control groups, they're own control
 2. More efficient, require fewer patients
- Disadvantages
 1. If one group has a curative treatment, then patients cross-over out of that group, raises ethical concerns
 2. Only useful for chronic, stable illnesses where treatments aren't "cures"
 3. If drugs have long half-life, require a long "wash-out" period during cross-over

Cross-Over Design

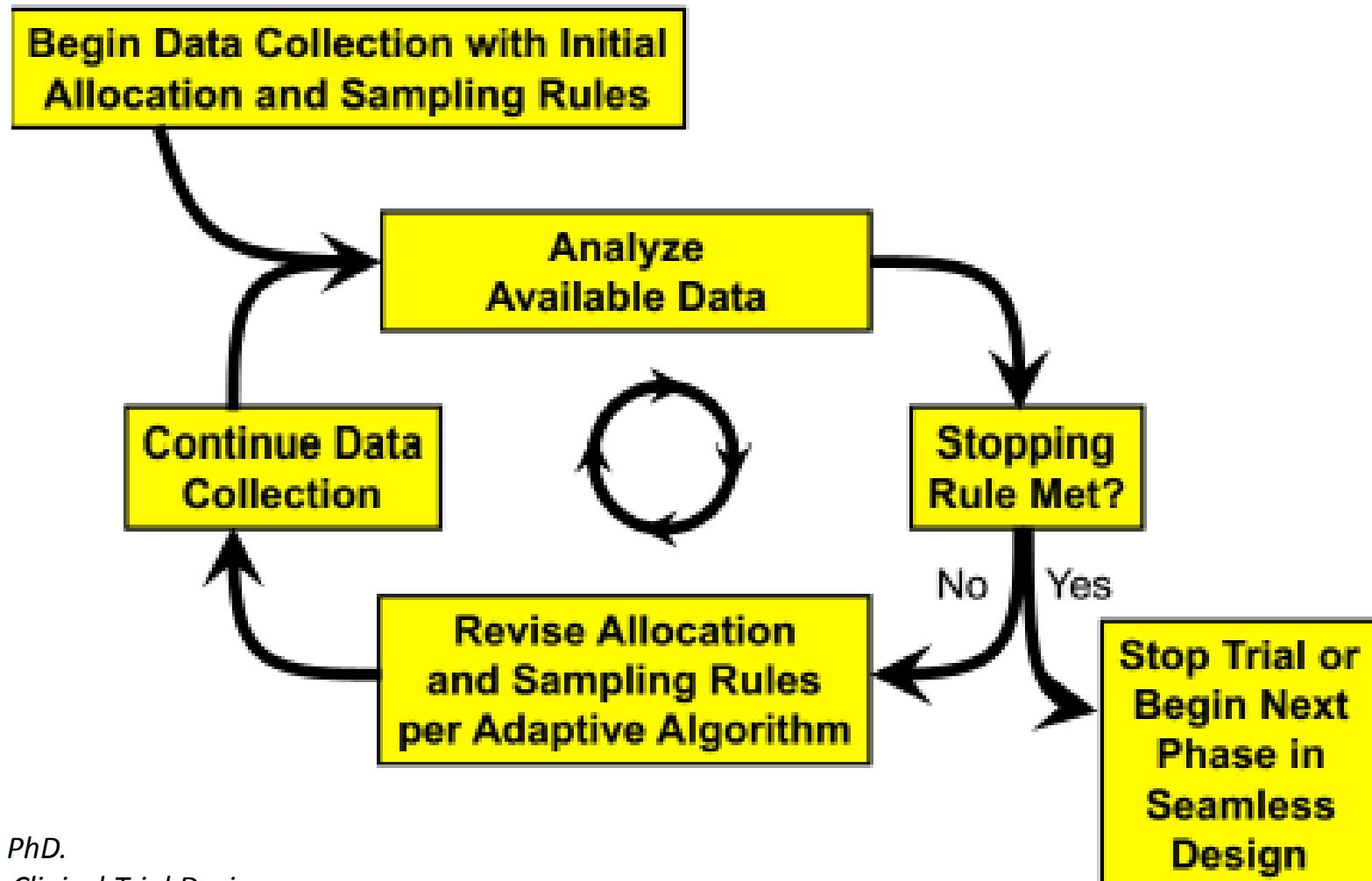
- Need to determine:
 - Sequence of groups
 - The period of time each treatment group is studied
 - Determine longitudinal analyses and how much time needed to properly assess those biomarker(s)
 - Length of wash-out period
 - Based on drugs being studied
- Most common design is a 2x2 cross-over: 2 drugs (A and B)
 - Need to determine which drug is better

	Period 1	Period 2
Sequence AB	Drug A	Drug B
Sequence BA	Drug B	Drug A

Adaptive Design

- Uses accumulating data to decide how to modify aspects of the study as it continues, without undermining validity and integrity of trial
 - Latter portion of the clinical trial is adapted/modified based on earlier portion of trial
- Goal is to learn from accumulating data and to apply what is learned as quickly as possible
- Shorten time of drug development from separate phase II and phase III trials and combine into one adaptive design trial
- Design includes a prospectively planned opportunity to modify one or more aspects of study design and hypothesis
 - Based solely on interim analyses of data from subjects in that trial

Adaptive Design



Presentation by Lewis, Roger, MD, PhD.

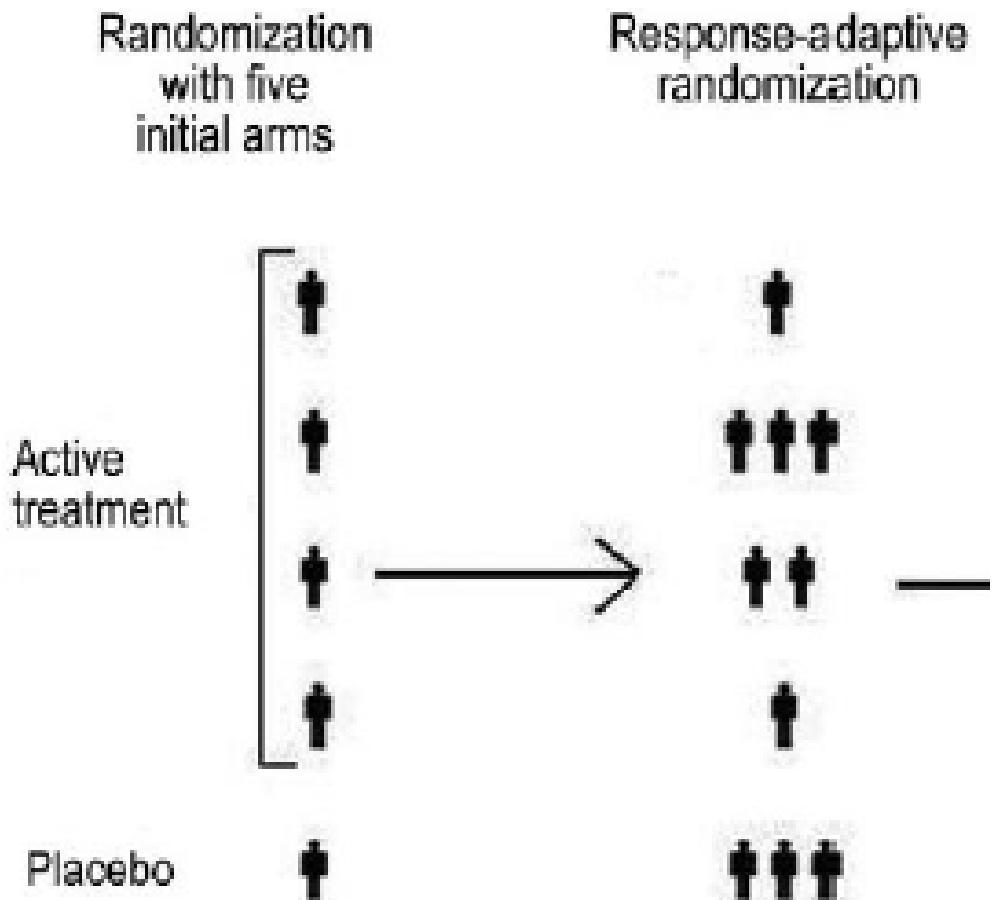
An Overview of Bayesian Adaptive Clinical Trial Design

Types of Adaptive Design Trials

- Adaptive randomization
 - Adaptive modifications of treatment randomization probabilities
 - Change the randomization algorithm
- Adaptive dose-finding
 - Find the promising doses to carry forward to phase 3
- Sample Size re-estimation
- Adaptive seamless-design

Adaptive Randomization Design

- Unequal probabilities of treatment assignment both prospectively and after review of response to previously assigned subjects
- Assign more subjects to promising treatments to increase trial success
- Response adaptive randomization



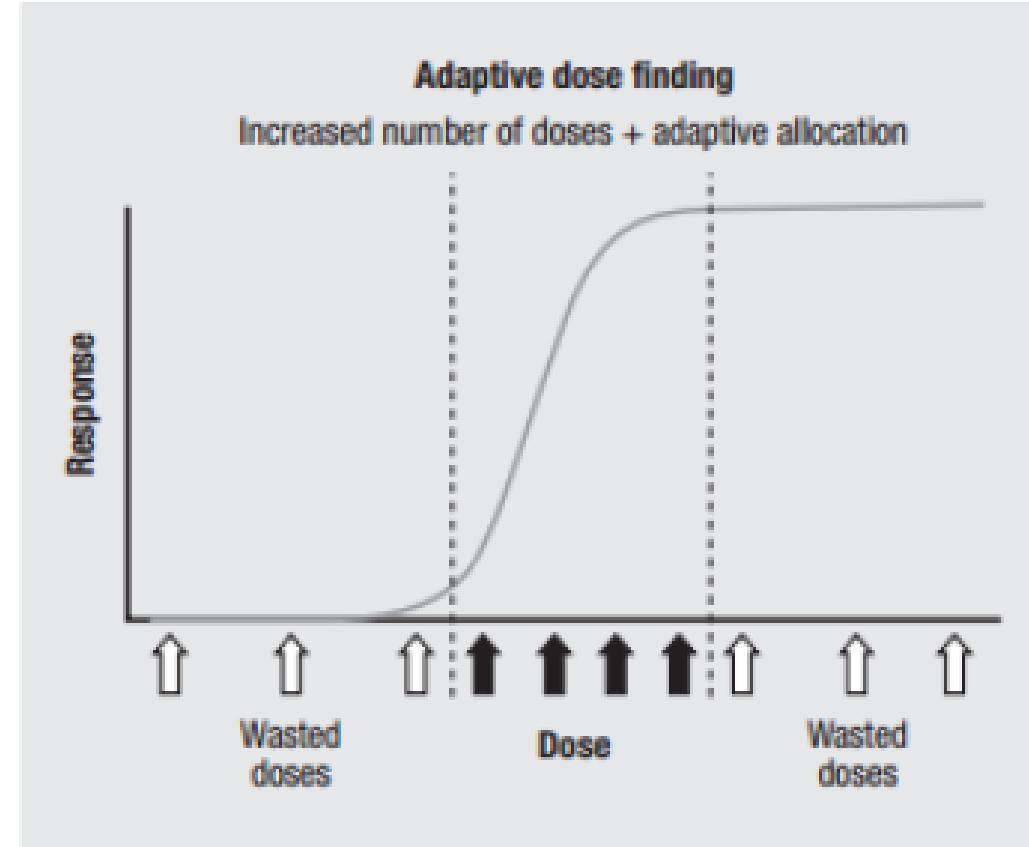
Adaptive Dose Finding Design

- **Traditional Dose Finding**

- Select a few doses for a phase 3, but those doses may not produce a dose-response curve
- Subjects may be exposed to sub- or supra-therapeutic doses unnecessarily

- **Adaptive Dose Finding**

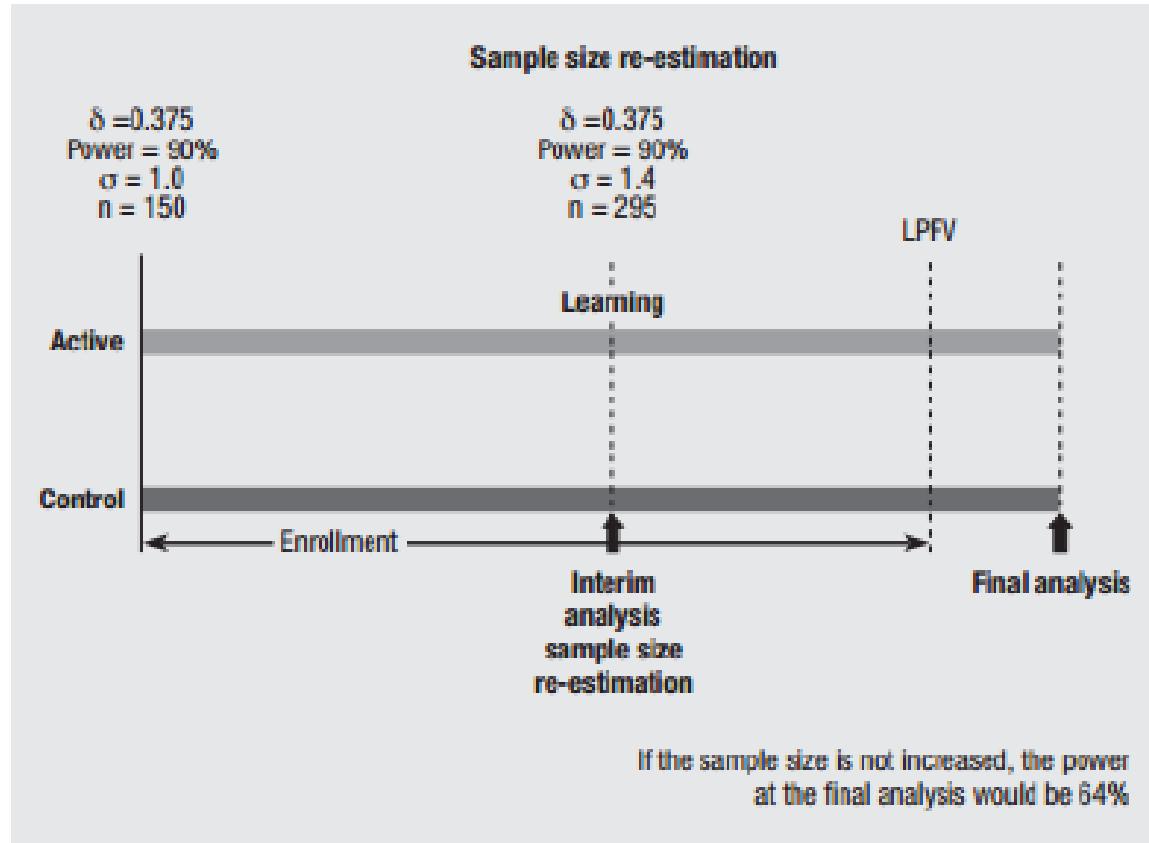
- Few subjects allocated to wide range of doses to explore dose-response
- Adaptive allocation/randomization based on response



Orloff, John, Frank Douglas, Jose Pinheiro, Susan Levinson, Michael Branson, Pravin Chaturvedi, Ene Ette, et al. "The Future of Drug Development: Advancing Clinical Trial Design." *Nature Reviews Drug Discovery* 8, no. 12 (December 2009): 949–57. doi:10.1038/nrd3025.

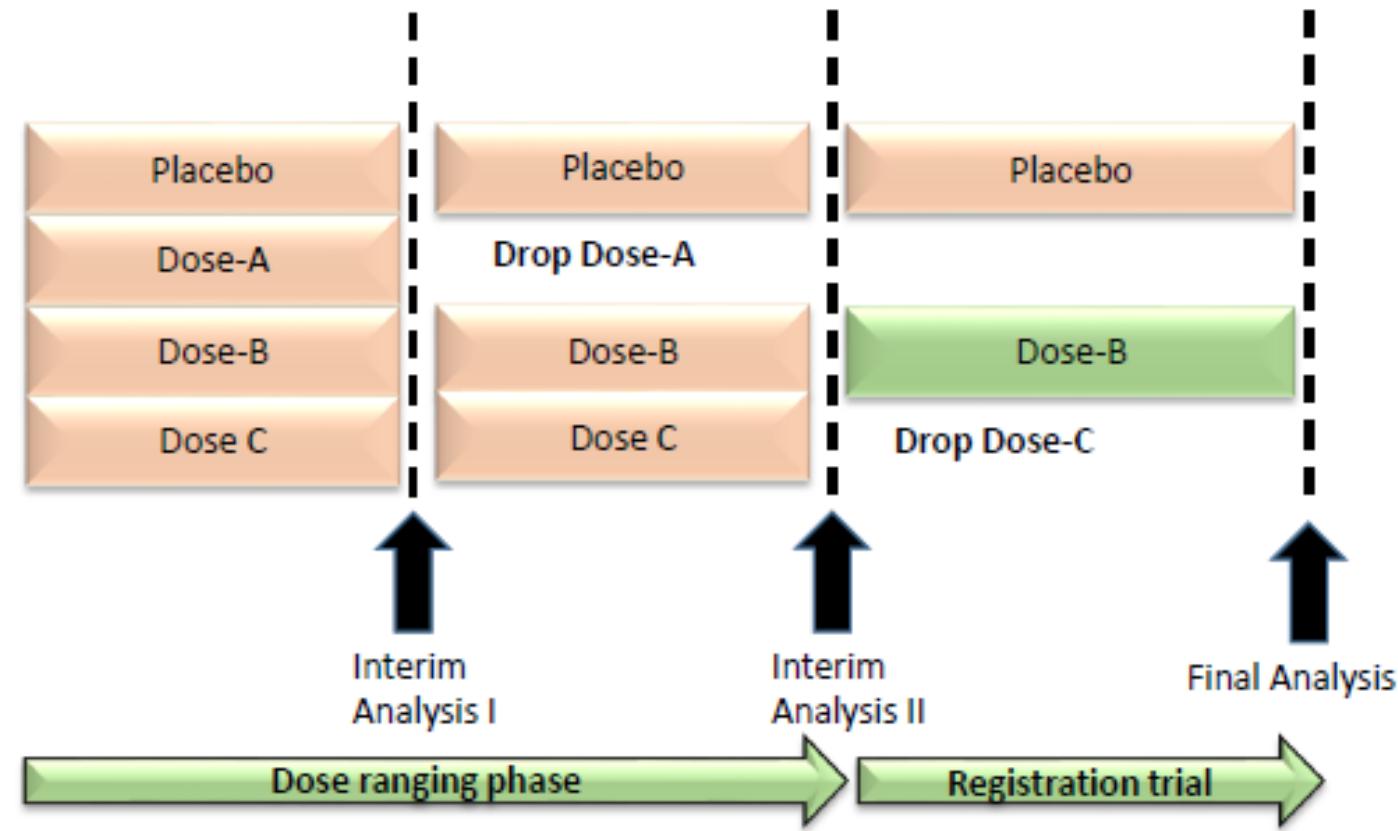
Sample Size Re-Estimation

- Beneficial to maintain statistical power to determine a difference
- Example:



Adaptive Seamless Phase II/III Design

- Phase II portion: a dose ranging, dose-finding phase
- Can use the adaptive dose-finding design within this phase
 - Drop sub- or supra-therapeutic doses
 - Add another dose
- Once optimal dose selected based on dose/response, perform phase III phase without having to close trial, design a separate phase III, wait for IRB, etc...



Disadvantages to Adaptive Design Trials

- Requires elaborate planning, simulations and logistical considerations
 - Compared to traditional trial design
- Requires collaborative effort of many different sections of an organization
- Internal resistant and aversion to change due to the flexible nature of adaptive designs
- Needs greater awareness among regulators and sponsors

Phase 3 Design

- Objectives: confirm efficacy and safety with a ***totality of evidence***
 - Large enrollment ($n > 100s-1000s$)

Adequate and Well-Controlled Trials

- Clear statement of the objectives
- Clear summary of analysis methods
- Valid design that permits comparison
- Adequate patient selection
- Minimum bias: trial, investigators
- Well-defined and reliable outcomes
- Valid statistical analysis methods

Randomization

- Assures that subject populations are similar in both test and control groups
- Avoids systematic differences between groups with respect to known or unknown baseline variables that could affect outcome
- Provides a sound basis for statistical inference
- Without randomization, no ability to eliminate systematic differences between treatment groups – a major problem of many studies

Types of Randomization

- Intent-to-Treat (ITT): a comparison of treatment groups that includes all patients as originally allocated after randomization
 - Does not account for whether they dropped out early, or never actually started treatment
- Per-protocol (PP) analysis: a comparison of treatment groups that only includes those patients who *completed* the treatment they were originally randomized to
 - This can lead to bias, since “healthier” patients at the start more *likely* to finish therapy, therefore, would have better response values

ITT

- Consistent with randomization
- Preserves unbiased testing
- Reflects “real-life” scenarios
 - Non-adherence to treatment schedule by patients is part of the outcome

PP

- Answers the question: Is the drug efficacious in patients who actually took the drug as prescribed?
 - compared to taking the drug *as intended* (ITT)
- Therefore, provides an accurate estimate of drug effect

Example: ITT vs PP

- n=501 patients recruited to a disk herniation trial to assess whether surgery was helpful
 - n=245 randomized to surgery
 - n=256 randomized to no surgery
- Of the n=245 assigned to surgery, 40% (n=105) never had surgery
- Of the n=256 assigned to no surgery, 45% (n=116) did have surgery
 - extenuating circumstances may have convinced PI to switch that patient in their best interest
 - called “crossing over”

	Bodily Pain (+ve change indicates surgery beneficial)	Physical Function (+ve change indicates surgery beneficial)	Disability (-ve change indicates surgery beneficial)
ITT analysis (N=501)	2.8 (-2.3 to 7.8)	1.2 (-4.1 to 6.5)	-3.2 (-7.8 to 1.3)
Per Protocol (N=280)	15.0 (10.9 to 19.2)*	17.5 (13.6 to 21.5)*	15.0 (18.3 to 11.7)*

Blinding

- Avoids placebo effect
 - If subject knows they have test drug, they may report more favorable outcomes because they expect a benefit or might be more likely to stay in a study if they knew they were on the test (active) drug
- Knowledge of treatment assignment could:
 - Affect the vigor with which on-study or follow-up data is obtained
 - Affect decisions about whether a subject should remain on treatment or receive concomitant meds
 - Affect decisions regarding inclusion of a given subject's results
 - Affect choice of statistical analysis

Types of Adequate/Well-Controlled Trials

- Placebo concurrent control
- Dose-response concurrent control
- No treatment concurrent control
- Active treatment concurrent control
- Historical control

Placebo Controlled Trial

Advantages

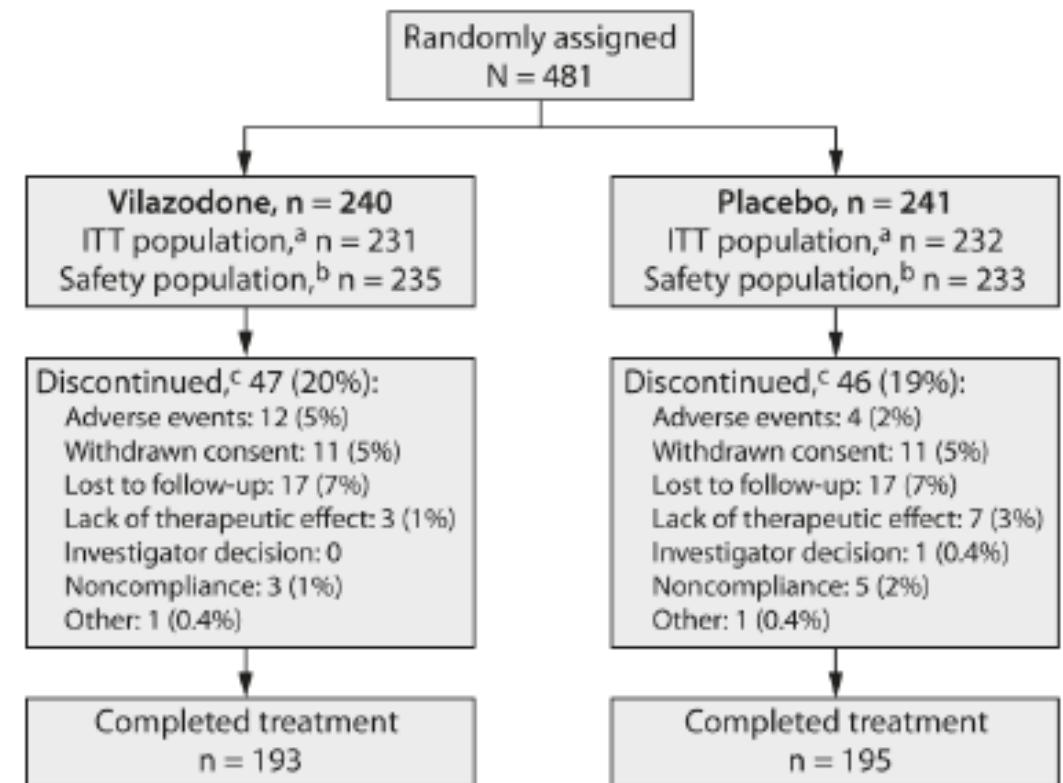
- Ability to demonstrate efficacy
- Measures *absolute* efficacy and safety
- Efficient
- Minimizes effect of subject and investigator expectations

Disadvantages

- Ethical concerns
- Patient and physician practical concerns
- Generalizability
- No comparative information

Placebo Controlled Trial

- Example of a randomized, placebo-controlled trial design



^aRandomly assigned patients who received at least 1 dose of study medication and who had at least 1 postbaseline efficacy assessment.

^bAll randomly assigned patients who received at least 1 dose of study drug.

^cN (%) of randomly assigned patients in each group.
Abbreviation: ITT = intent to treat.

Dose Controlled Trial

Advantages

- Similar to placebo-controlled
- Efficient
- Possible ethical advantage

Disadvantages

- If no pair-wise comparisons built in, then may be difficult to determine optimal dose
 - Assuming there's a positive correlation b/w dose and efficacy
- Not uncommon to have no dose/response over a dose range
 - If no placebo group, then this design is uninformative
- If therapeutic range unknown prior, designed dose range may be sub- or supra-therapeutic
- Design may be less efficient than placebo-controlled for showing a drug effect
 - BUT, if designed well, dose-controlled trials can provide better dose/response info

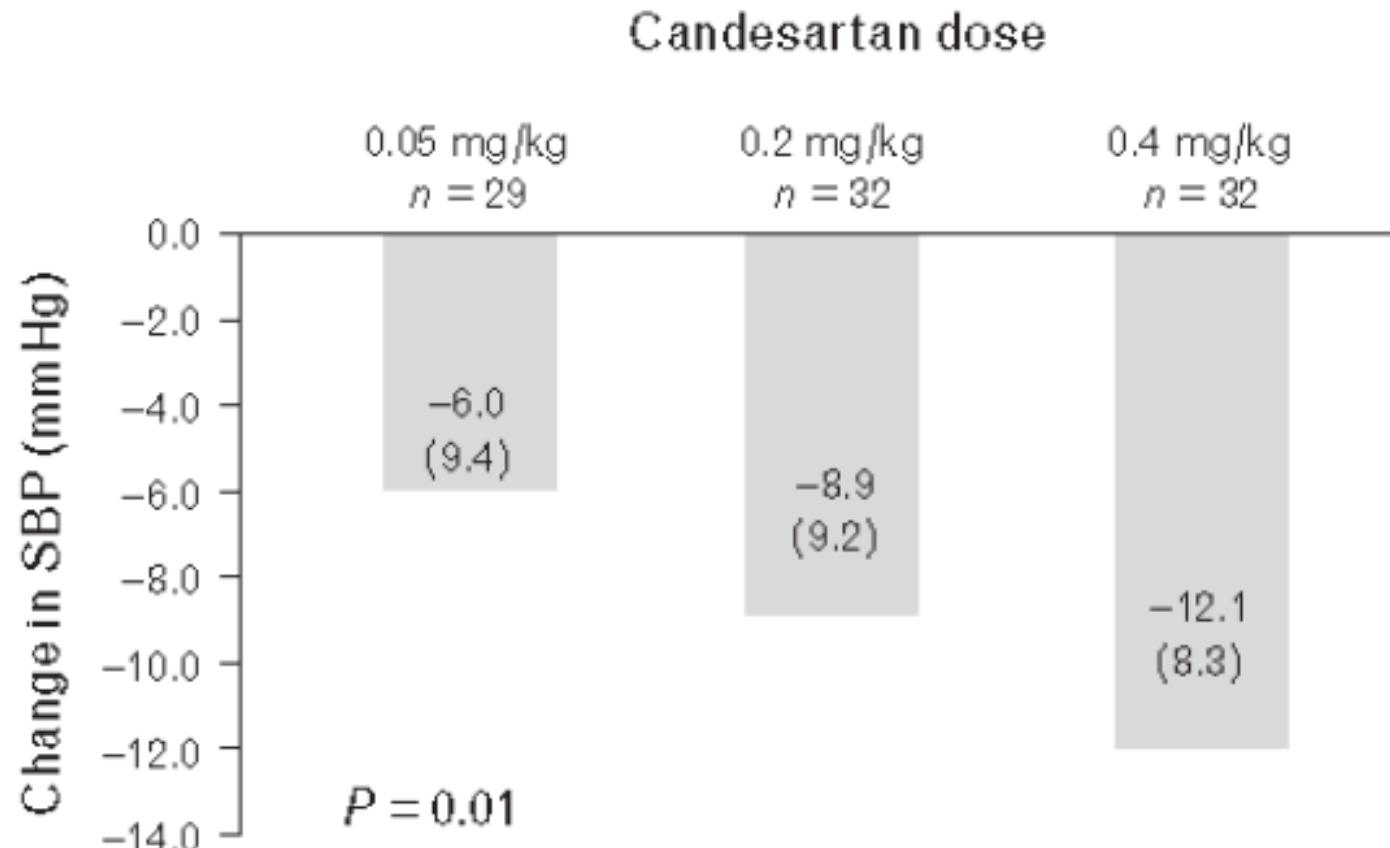
Dose-Controlled Trial

Example:

- Candesartan efficacy, safety tested in a 4-week, randomized, double-blind (both investigators and subjects), dose-ranging study
 - Followed by a 1-yr open label (unblended) treatment phase
 - Doses: 0.05, 0.2, or 0.4 mg/kg/day
 - Given as a liquid suspension
 - Subjects: n=93 children aged 1-5 ys, 74/93 had underlying renal disorders
 - Biomarker for primary efficacy measure was change from baseline seated systolic blood pressure (SBP)

Dose Controlled Trial

- Monitor response in seated SBP at a pre-defined time point in the trial from the seated SBP at baseline (pre-drug)
 - Provides the “change” in SBP
- Perform for each dose level
- Look for a dose-response trend



Active Controlled Trial

Advantages

- Ethical and practical advantages
- Comparative information

Disadvantages

- Trial size
- Assay sensitivity
 - Ability to distinguish an effective treatment from a less effective or ineffective treatment
 - Different implications for trials intending to show differences between treatment (superiority) and trials intended to show non-inferiority

Active Controlled Trials

- Most active controlled trials are non-inferiority, equivalence trials
- Designed to establish efficacy of a new treatment
- Compares new treatment (test) with a known effective treatment (control)
- Requires pre-existing knowledge of the response variable and to what extent a change is needed (delta response) to be “effective”
 - This delta response, or simply “delta” is the margin of equivalence, or non-inferiority, that the test drug will have to induce to be considered “equivalent” or not worse than (non-inferiority) the control
- Non-inferiority delta/margin cannot be greater than the *smallest effect size that the active control would be reliably expected to have*

Historical Controlled Trial

Uses an external control (historical data) to compare efficacy and safety

Advantages

- All patients/subjects can receive active drug
 - No placebo group
 - More attractive to subjects
 - Important for test drugs in rare diseases

Disadvantages

- Cannot be blinded
 - Potential for bias from subjects and investigators
 - CAN blind post-hoc analysts
 - Useful **only** when impossible to blind treatment
 - Tablet/capsule itself is unique and can be ID'd easily
 - Toxicities easily recognized

Historical Controlled Trial

Example: Gleevac® (Imatinib)

- Phase 3 trial in patients with chronic myeloid leukemia (CML)
 - 3 types of CML: chronic, accelerated, blast
 - Single arm, 400mg dose
 - Patients w/ CML who failed first-line treatment
 - Had no comparative survival data or sustainability of response
- A candidate for accelerated, but not full, approval

Summary of Phase 3 Controlled Pivotal Trials

Trial Objective	Placebo	Active non-inferiority	Active Superiority	Dose Response (D/R)	Placebo + Active	Placebo + D/R	Active + D/R	Placebo + Active + D/R
Measure Absolute effect size	Y	N	N	N	Y	Y	N	Y
Show existence of effect	Y	P	Y	Y	Y	Y	Y	Y
Show Dose-Response relationship	N	N	N	Y	N	Y	Y	Y
Compare therapies	N	P	Y	N	Y	N	P	Y

Totality of Evidence

- FDA typically requires two adequate, well-controlled studies to show efficacy that is convincing individually within each trial
 - Taken together, trials demonstrate a totality of evidence of effectiveness and safety
- To best ensure “efficacy”, sponsors must wisely choose an outcome/response that itself can ensure the effect is *clinically meaningful*
 - Outcome responses chosen *a priori*

Phase 4

- Post-marketing trial
 - After FDA approve for sale in USA
 - If unexpected toxicity (adverse events) observed with wide marketed use, FDA can request a Phase 4 safety trial
-
- Possible that FDA could revoke the drug's approval for use in USA
 - Sponsor could also withdrawal drug from market
 - Example: Vioxx®, a COX-2 inhibitor, NSAID approved in May 1999
 - Merck was conducting long-term safety in patients with recurrent colon polyps
 - Noticed there was a small exposure/response with increased risk for heart attack and stroke, particularly for patients on therapy for 18+ months
 - Merck withdrew Vioxx® in September 2004 and erupted a scandalous account of deceiving the FDA and the public

Use of Pharmacometrics

- In the case of Vioxx®, Merck had several years and trials' worth of safety data *suggesting* an increased risk of cardiovascular events (CE)
 - Heart attack, stroke
- Without actually knowing, and just assuming..... If pharmacometrics could have been properly utilized to identify this risk earlier:
 - An exposure/response analyses could have been modeled with Vioxx® AUC and risk of cardiovascular event (odds ratio or hazards ratio)
 - AUC can subsequently be predicted from dose to get a dose/response analysis
 - Pharmacometrists could have determined a dose that induced an “acceptable” level of risk
 - Effectively communicate with executives
 - Simulate trials to choose an optimal dose that is effective without increased risk of CE

Day 1 Summary

- Today we were introduced to drug development process
- Understand how exposure/response analyses are a crucial focal point in determining success/failure of a drug candidate
- Learned how to calculate exposure parameters via NCA PK
- Understanding of clinical development of drug candidates to better understand how to model subsequent trial data
- Tomorrow, we begin exposure response modeling
 - Four different sections to cover over next three days

Day 2

9:00 - 10:15am:

- Day 1 recap, overflow
- Understanding clinical trial data w/ statistics

10:15-10:30am:

- Break

10:30-12pm:

- Exposure/Response modeling I
 - Linear models
 - Continuous vs Continuous (linear regression)

12:00 – 1:00pm:

- Lunch break

1:00 – 2:30pm:

- Exposure/Response modeling II-A

2:30 - 2:45pm:

- Break

2:45 – 4:15pm:

- Exposure/Response modeling II-B
 - Logistic regression
 - Bernoulli Distributions
 - GLM

Exposure/Response Analyses

- How to measure exposure (Pharmacokinetics; PK)
- How to measure response (Pharmacodynamics; PD)
 - Several aspects of clinical trials that can be measured or monitored for a biological response to the drug.
 - Quantifying these biological responses via biomarkers
 - Can then correlate w/ quantifiable PK exposures
 - Need basic understanding of statistics, differential calculus, and pharmacology

Exposure Analysis

- Quantifying the amount of drug exposure a subject receives after being administered a given dose, in a given route, over a given frequency
- Requires pharmacokinetic (i.e. drug movement) analyses
- Exposure is a **dependent** variable
 - Dependent upon independent variables: age, weight, dose, organ function, etc.
- Need to measure/quantify drug concentration in biological matrix (typically plasma) at several time points post dose
 - Get *estimates* of drug absorption, distribution to various locations in body, metabolism and elimination (ADME)
 - Every parameter *estimate* is just that....meaning there is some degree of *uncertainty* associated with that estimate. Quantification of that uncertainty is standard error
 - Use SE to calculate a confidence interval (CI)

Response Analyses

- Need to know if drug is inducing a response
 - Based on quantifiable response endpoint (biomarker) in both treatment and placebo arms of a clinical trial
 - Statistical analyses of whether drug significantly better than placebo based on null hypothesis

Understanding Clinical Data

- Measurement
 - How we obtain data
 - Example: blood pressure, body weight, height, serum creatinine, LFTs
 - Because measurements are made using assays, there is always assay variability
 - A source of error, typically an acceptable level of error allowed
 - Example: body weight a function of scale used. If step on 20 scales, you'll measure 20 different body weights. What is the “true” body weight? Impossible to know, but acknowledge some level of error
- Observation
 - The unit (e.g. the subjects in a trial) upon which measurements are made

Observations

- **Population** – consists of **all** subjects of interest
 - Characterized by “parameters”
 - Target population
 - For a prostate cancer drug, all men w/ prostate cancer
 - For a blood pressure drug, all persons with high blood pressure
 - Study population
 - For a given clinical trial of a prostate cancer drug, all men in USA with prostate cancer
 - It is ***impossible*** to conduct a trial in an entire population...
- **Sample** – a (hopefully) unbiased randomized subset (N) of a population
 - Characterized by “statistics”
 - Sample chosen to draw conclusions about the entire population
 - Statistical inference

Types of Clinical Data Variables

- **Qualitative or Categorical** – a variable with categories
 - Ordinal – meaningful ranking or scale, but not quantified
 - Example: pain status (mild, moderate, severe)
 - Example: cancer staging or grades (1, 2, 3, or 4)
 - Nominal – NO meaningful rankings
 - Binomial – yes/no; gender (male/female), genotype (for some genes)
 - Other – blood types, race, genotype (for some genes)
- **Quantitative** – a variable with numeric values
 - Continuous – measured on a continuous scale
 - Example: age, body weight, drug concentration (i.e. drug exposure)
 - Discrete – measured on a discrete scale
 - Example: number of seizures within a time period (count data), survival (event data)

More on variables....

- **Dependent** – a measured variable dependent on experimental conditions (i.e. independent variables)
 - Aka the **response** variable
 - Example: plasma concentration, blood glucose, blood pressure, etc
- **Independent** – a variable that is set
 - Aka the **predictor** variable
 - Example: body weight, age, dose, etc

Exposure/Response Regulation

- FDA doesn't explicitly outline how they regulate exposure/response analyses (aka dose/response)
- One clause in Food, Drug, and Cosmetic Act (314.126) provides *guidance* on how to conduct and analyze dose/response
 - Clause 314.126 states a dose/response study has 2 purposes:
 1. An adequate, well-controlled study can show efficacy using pair-wise vs placebo or another dose (dose/response) and a documented increased slope of effect vs dose
 2. Provide a basis for dose selection, which needs additional criteria beyond just showing dose/response.

Standard Trial Design for Dose/Response

- Randomized, parallel, fixed-dose
- First used by Materson in 1978 for chlorthalidone for lowering BP
 - Showed that 25mg gave **same** effect as 200mg with MUCH LESS toxicity
- Now, randomized, parallel, fixed-dose studies are the norm for dose/response studies
 - In 1994, both FDA and International Conference on Harmonization of Technical Requirements for Registration of Pharmaceuticals for Human Use (ICH) *strongly encouraged* use of dose/response studies for drug development. And to know *SHAPE* of effects (effective and toxic)
 - No advice/guidance on HOW to use that data to select dose

Choosing a Dose

- Analyzing and describing an exposure or dose/response relationship doesn't tell you what dose to pick
- That depends on a few variables:
 - Urgency of need for effect
 - Degree of separation of efficacy and toxicity
 - Importance of toxicity (can you start w/ high dose then scale down?)
 - Extent of individual variability in PK and PD response to dose
 - Point of view
 - Some people argue for lowest dose with any effect, but not usually what's done)
 - Most dose/response curves show small differences between adjacent doses

Choosing a Dose

Choices:

1. High dose – if large separation between effective and toxic dose ranges, or if observed tox not dose-related

2. Low dose – if small separation between effective and toxic dose ranges.
 - Drugs in this category need titrated

Hypothesis Testing in Clinical Trials

- Trial designed to assess effects of a test drug vs placebo
 - Two-arm, parallel design
- Establish a null hypothesis (H_0) that effects caused by test is NO different than effects caused by placebo
 - H_0 always involves equality
- Effect usually refers to study's primary endpoint

Alternate Hypothesis (H_A)

- Less well-defined than the null hypothesis
- Only proved true once null is disproven

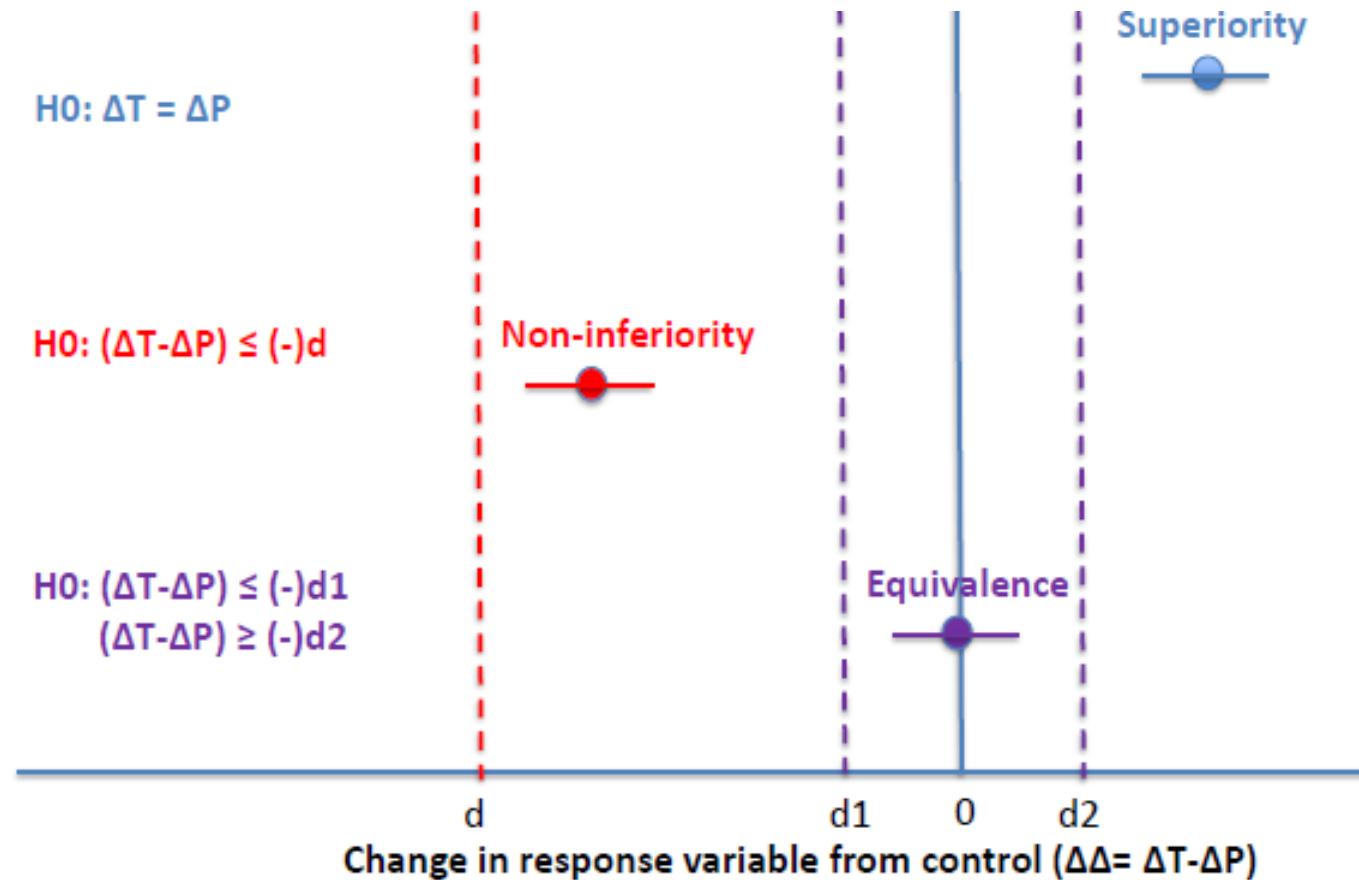
Types of Error

- Type I error(false positive; rejecting null hypothesis when actually true, i.e. assuming a real difference when there actually is no real difference)
- Type II (false negative; accepting the null hypothesis when actually false; i.e. assuming there's no difference when there actually is a real difference)

	Truth about Null hypothesis	
Decision based on Test Statistic	True	False
Fail to reject Null hypothesis	Correct	Type-II error (β)
Reject Null Hypothesis	Type-I error (α)	Correct (Power: $1-\beta$)

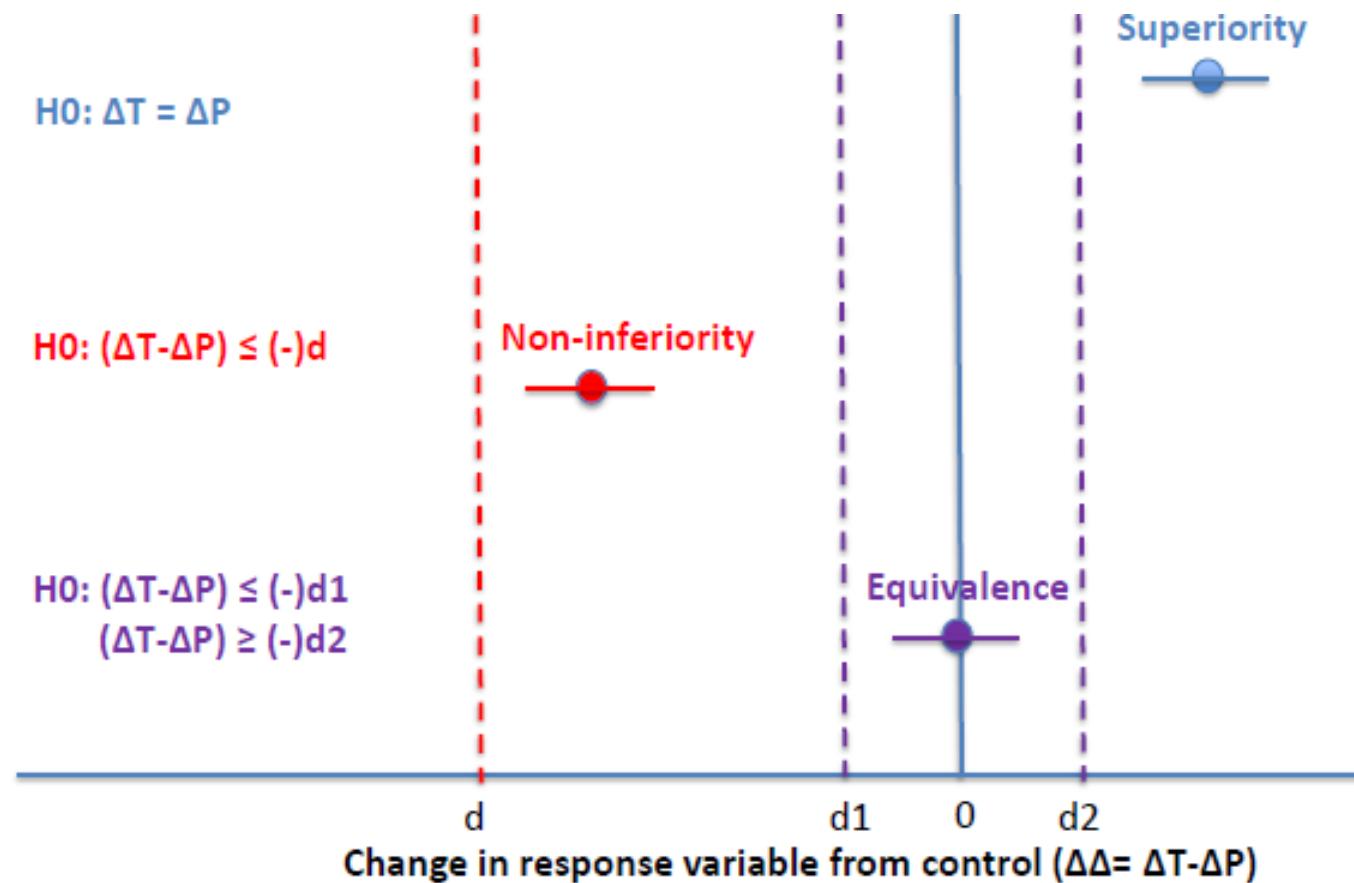
Hypothesis Testing in Clinical Trials

- Requires a good biomarker and assay
- Requires knowledge of efficacious changes in biomarker (“delta”)
- Understanding of basic statistics



Hypothesis Testing in Clinical Trials

- Determines the type of trial to design (superiority, equivalence, non-inferiority, etc)
- Determines what the null hypothesis (H_0) is...



Hypothesis Testing in Clinical Trials

H_0 (Superiority): change in biomarker from test drug is same as change in biomarker from placebo

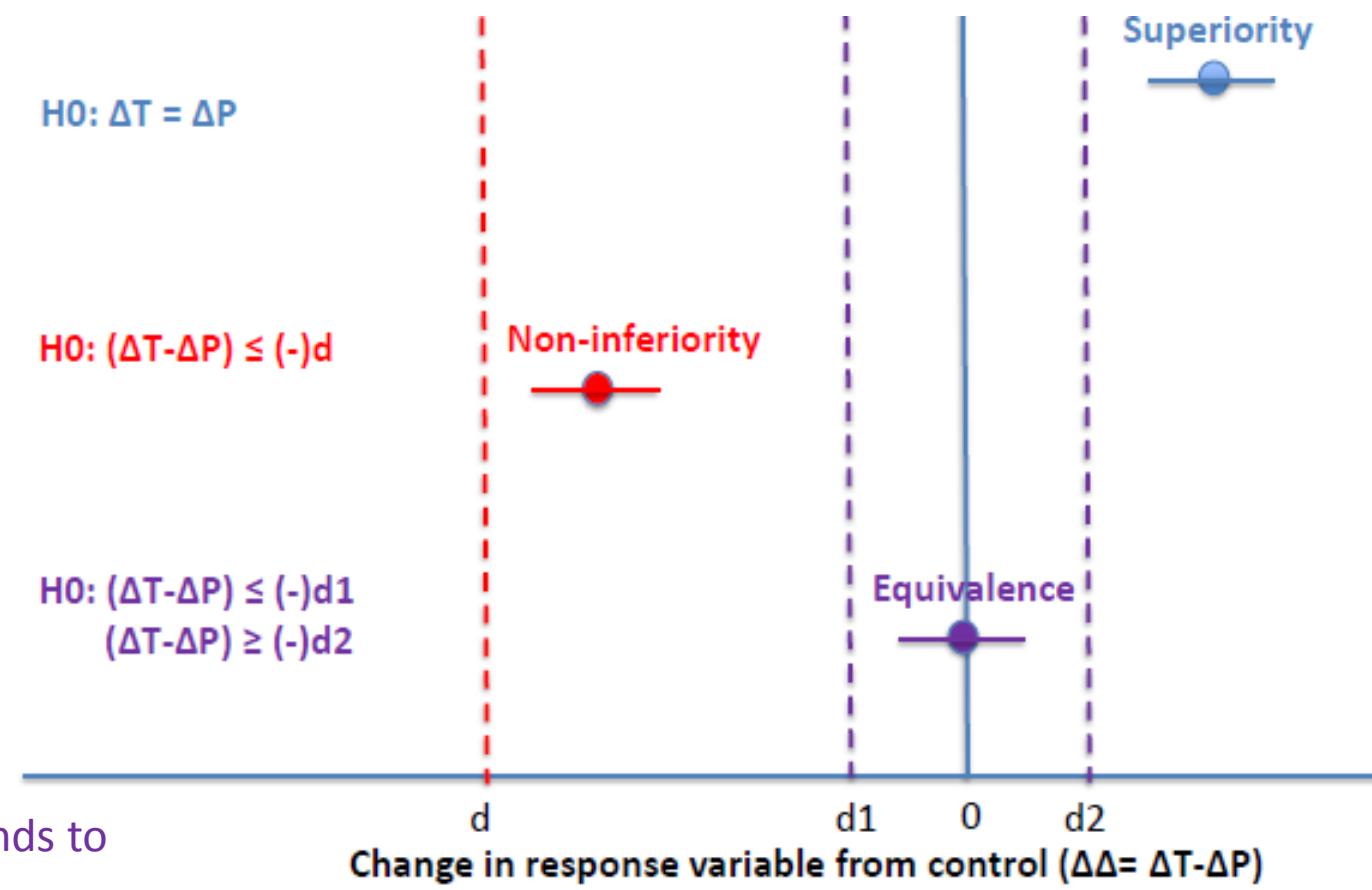
- Need clear evidence (factoring in some error) that test drug better than placebo to reject H_0 (accept superiority of test over placebo)

H_0 (Non-Inferiority): the difference between biomarker changes in test vs placebo is less than or equal to some pre-determined biomarker change from control

- Need clear evidence that difference in test-placebo is not worse than control to reject null (accept non-inferior)

H_0 (Equivalence): the difference between biomarker changes in test vs placebo is EITHER less than a pre-determined lower bound (d_1 ; usually 80% of "0") OR greater than a pre-determined upper bound (d_2 ; usually 125% of "0")

- Need clear evidence (factoring in some error) that Difference in test vs placebo is NOT outside these bounds to reject null (accept that test is equivalent to placebo)



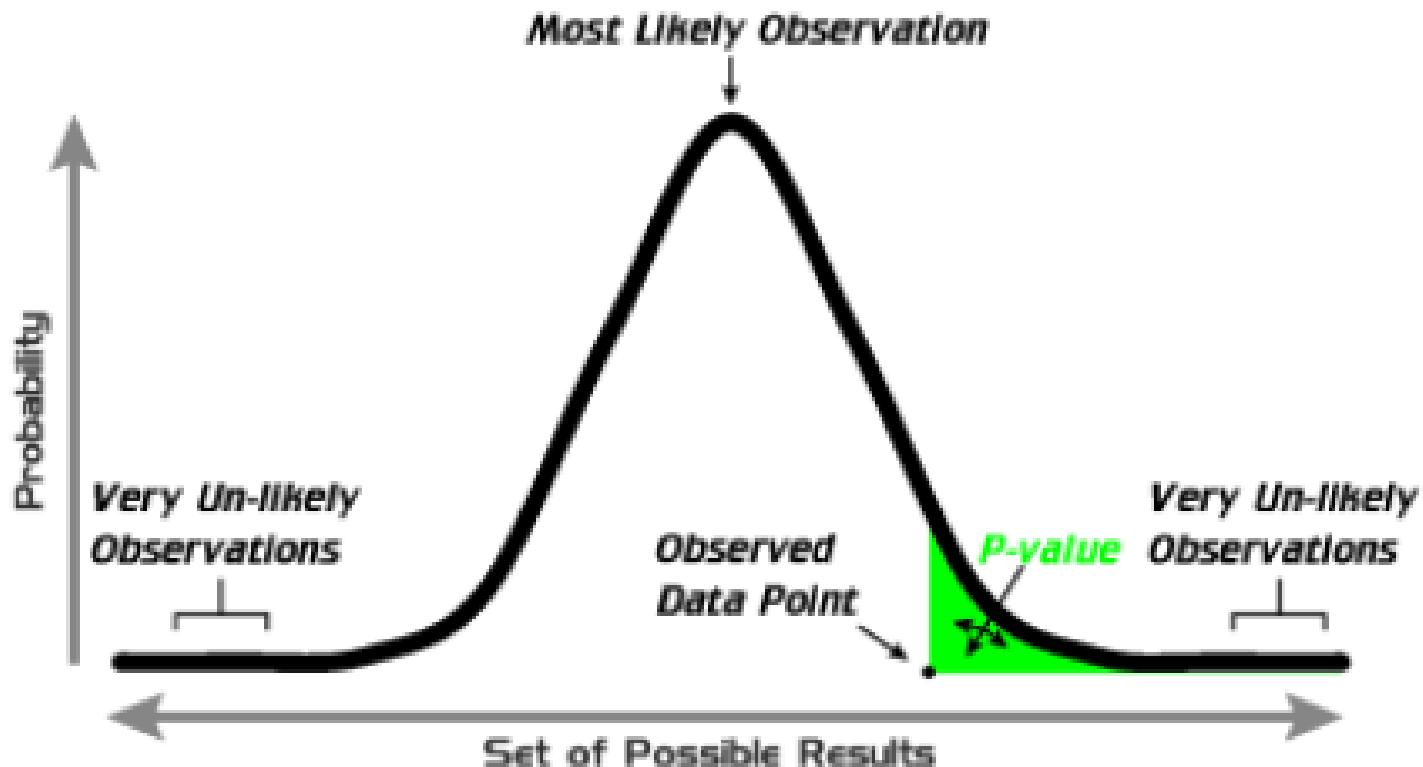
p-values in Hypothesis Testing

- p-values tell how likely the observed result is, referring to the probability that result obtained is by chance alone
- If $p=0.05$, null hypothesis has a 5% chance to be true due to error
- If $p>0.05$, there's a greater than 5% chance null hypothesis is true due to error, so cannot reasonably reject the null, must *accept*
- If $p<0.05$, there's a less than 5% chance null hypothesis is true due to error, so can reasonably *reject* the null hypothesis

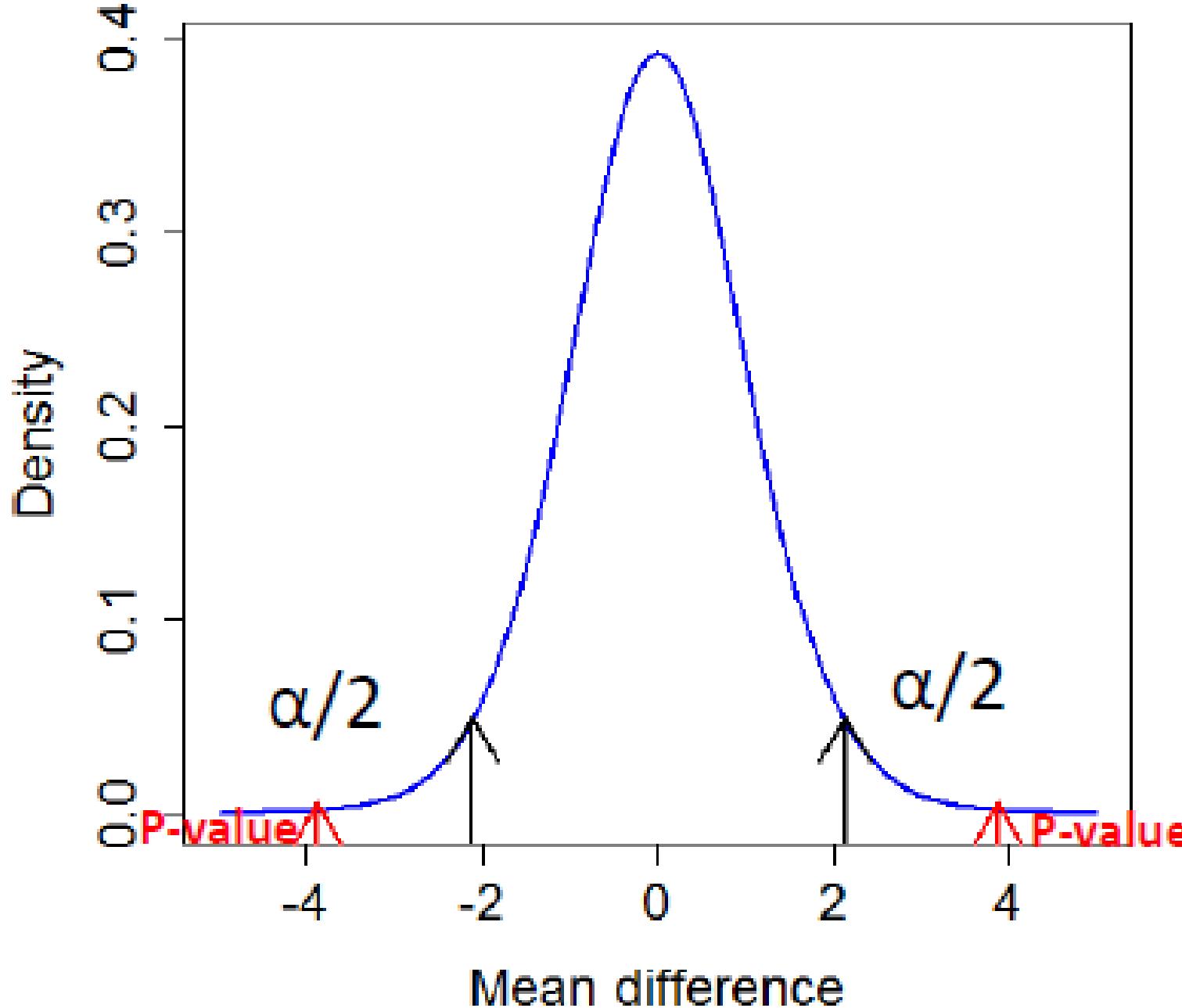
**assuming a type I error of 5% ($\alpha=0.05$)

p-values in Hypothesis Testing

- p-value refers to the **probability** of obtaining a test statistic at least as extreme as the one that was actually observed, assuming the null hypothesis is true



A **p-value** (shaded green area) is the probability of an observed (or more extreme) result arising by chance



t-distribution

- Central limit theorem states if you know population SD or Variance, then can use Z-statistic
 - However, this is usually impossible
 - Instead use t-statistic

Comparing Z- vs T-statistics

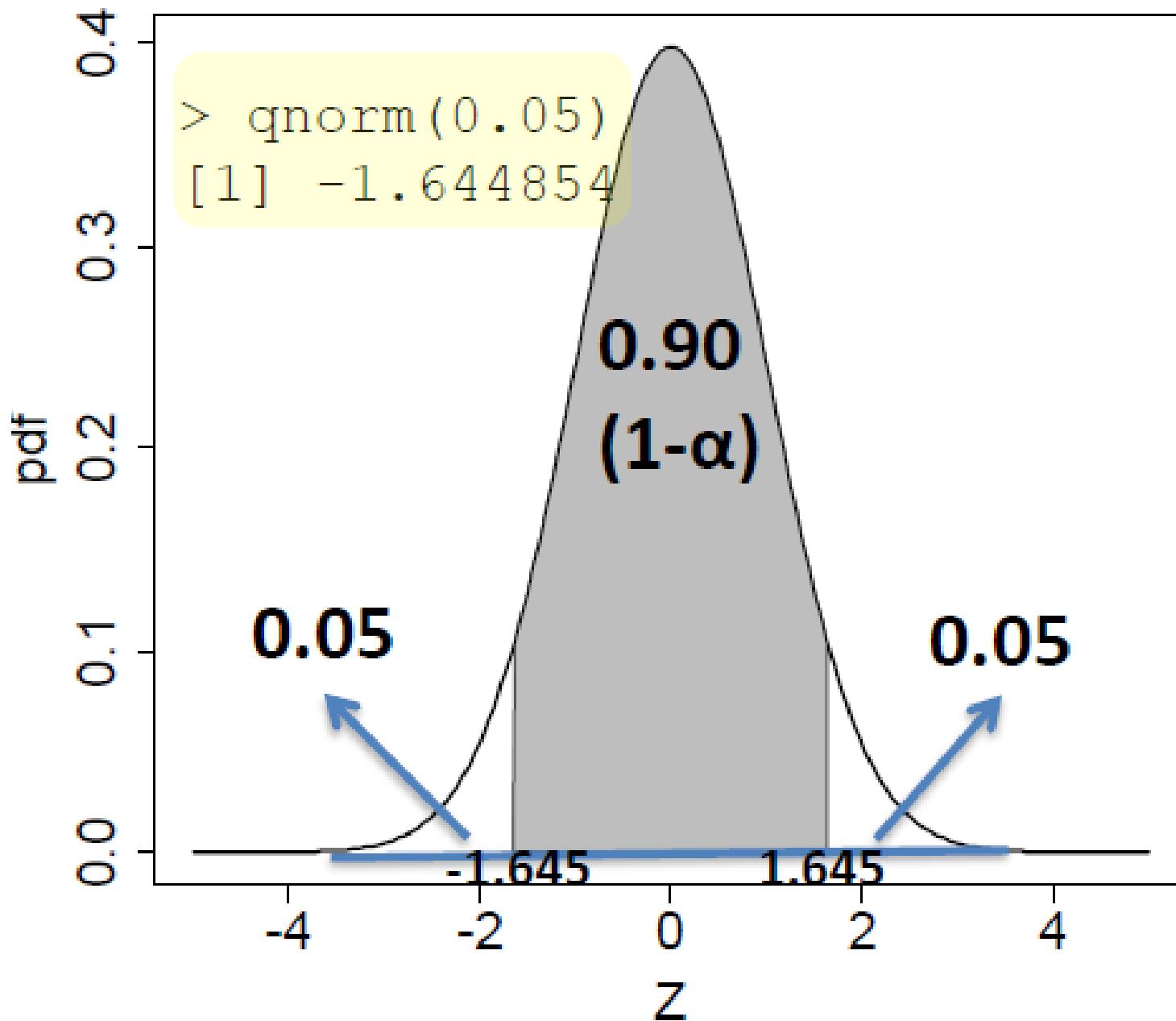
Confidence levels	Z-interval	T-interval
90%	115.80,123.16	115.62,123.35
95%	115.10,123.86	114.80,124.16
99%	113.72,125.24	113.09,125.88

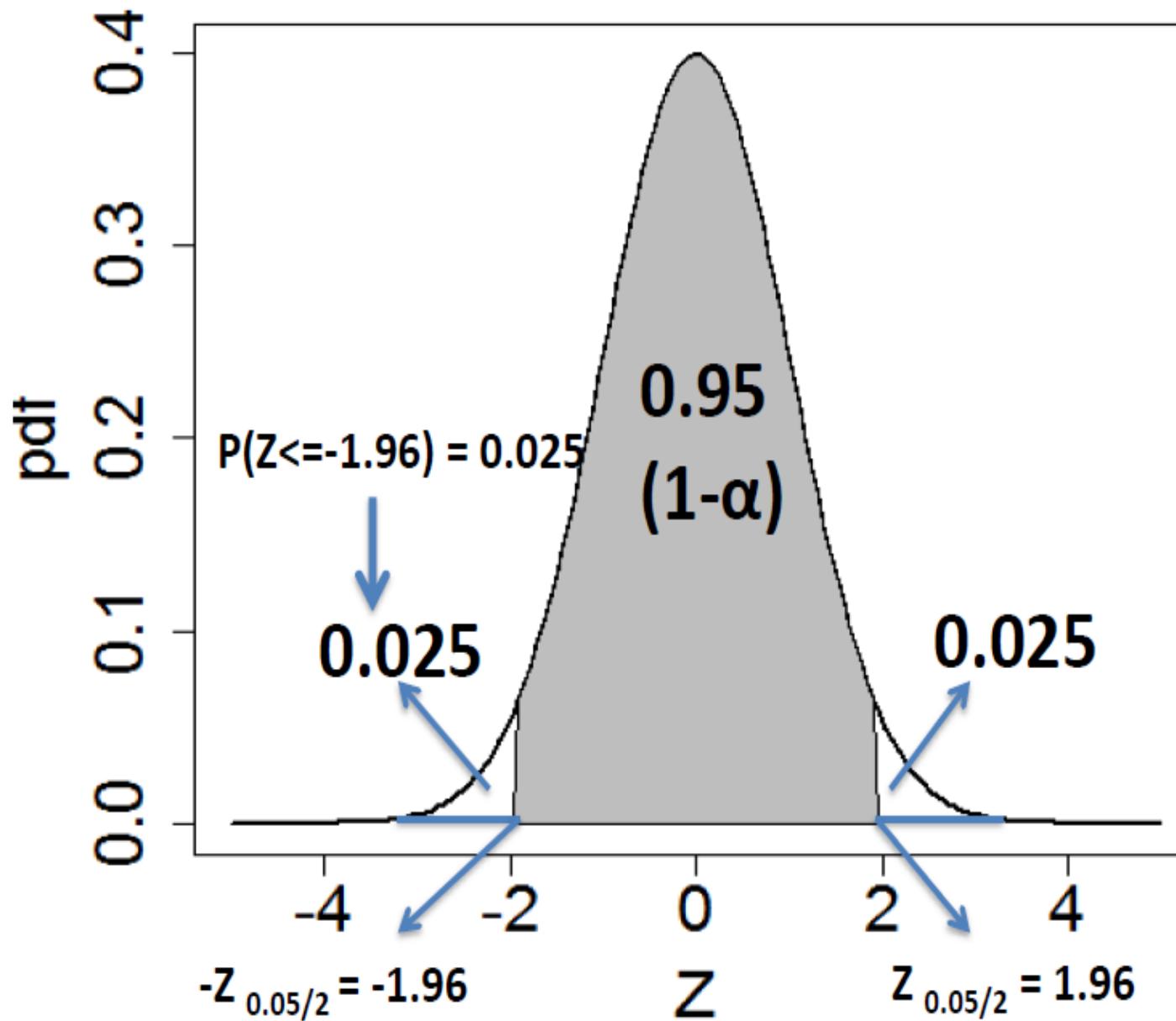
- Bottom line: Higher confidence wider the interval (both Z and t- interval).
- T-interval is wider than Z-interval in general.

Which interval do you use?

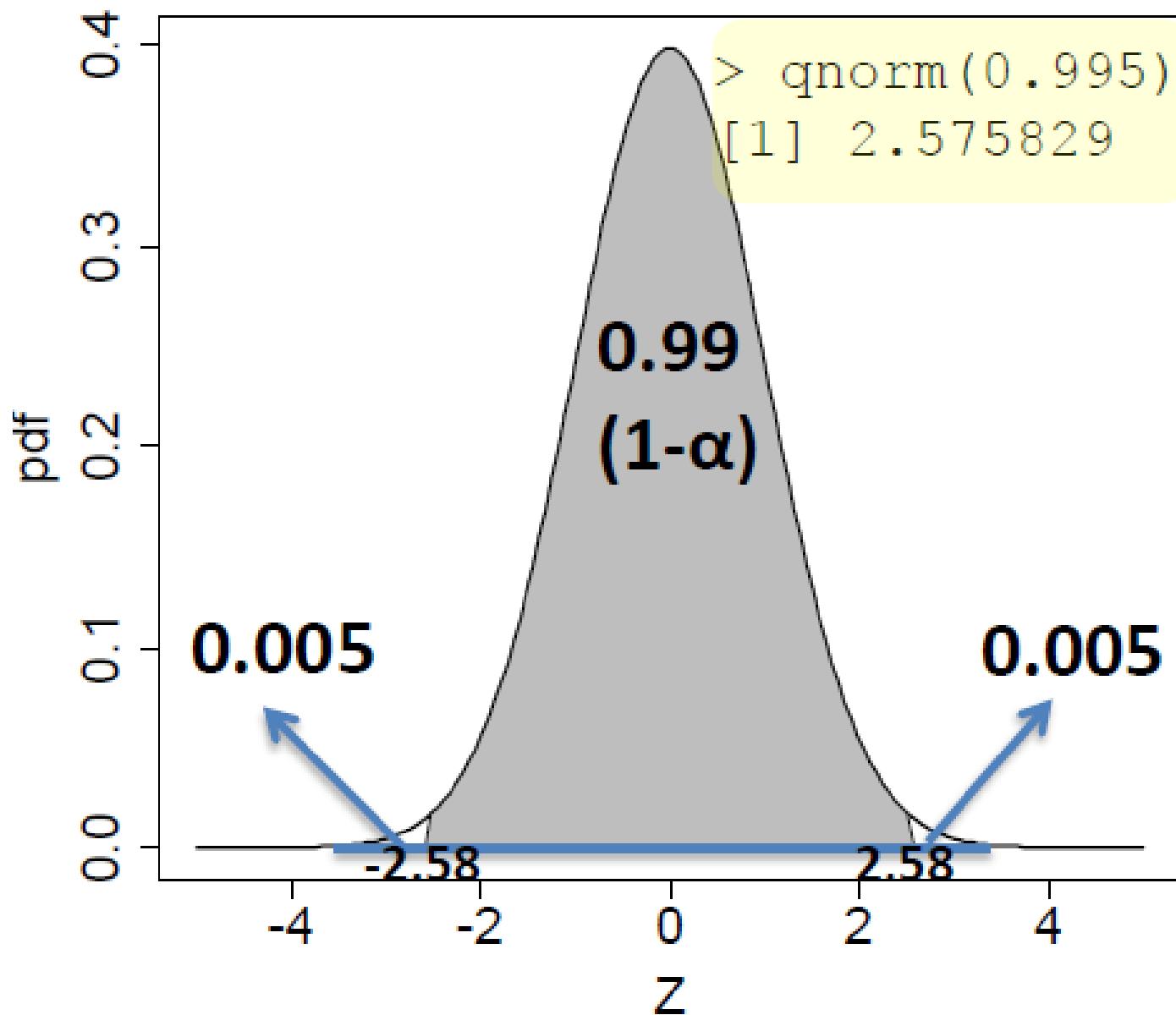
Condition	CI interval for μ to use
Yi's are Normal, μ is unknown, σ^2 is known	$\bar{y} \pm Z_{\alpha/2} \frac{\sigma}{\sqrt{n}}$
Yi's are Normal, μ and σ^2 are unknown (n<=30) : small sample size	$\bar{y} \pm t_{\alpha/2, n-1} \frac{s}{\sqrt{n}}$
Population distribution not known, μ and σ^2 are unknown (n>30) : large sample size Because of Central Limit theorem (only for sample mean)	$\bar{y} \pm Z_{\alpha/2} \frac{s}{\sqrt{n}}$

90%

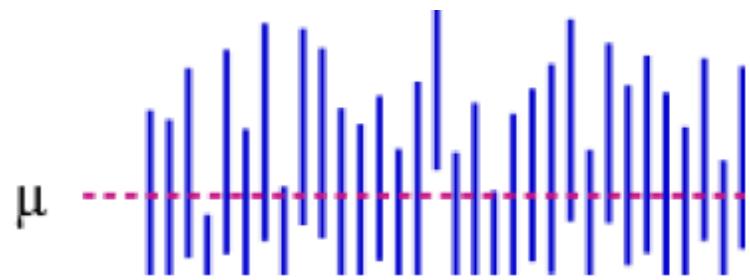




99%



Confidence Intervals



- CI consist of a range of values (an interval) that act as good estimates of the unknown population parameter
 - Based on data from a *sample* of the population
 - Impossible to collect data on the entire population, so must *estimate* the population parameter value (μ) based on distribution of a variety of samples
 - Because estimating population from sample, there's inherent error/assumptions
- Level of confidence of the CI would indicate the probability that the range captures the true population parameter given a distribution of samples
- 95% CI: means 95% of the observed confidence intervals will hold the true value of the population parameter

Confidence Intervals and Hypothesis Testing

- General Rule:
 1. If 95% CI does NOT contain 0, then reject the null at $\alpha=0.05$
 2. If 95% CI does contain 0, then do NOT reject the null at $\alpha=0.05$

Two-Sided Hypotheses

- When comparing treatments, null hypothesis states: test=placebo
- However, if reject the null, that could be due to:
 - test > placebo
 - test < placebo
- In this case, would perform a two-sided, or two-tailed, statistical test to compare the mean response of test vs mean response of placebo (+/- some error within each group mean)
 - A t-test (unpaired, since groups do not represent paired, repeated measurements)

One-Sided Hypothesis

- Relatively rare
- Only used when have previous data to suggest that test cannot be greater or less than placebo
 - Only one or the other, not both
- Example: antibiotics known to damage kidney cells, lowering renal function as measured by serum creatinine clearance
 - When antibiotics present, serum creatinine will ALWAYS be greater, NEVER lower, because there is a known physical mechanistic limitation
 - Either test antibiotic will increase serum creatinine, or induce no change
 - Never lower serum creatinine
 - In this case, a one-sided null hypothesis would be appropriate

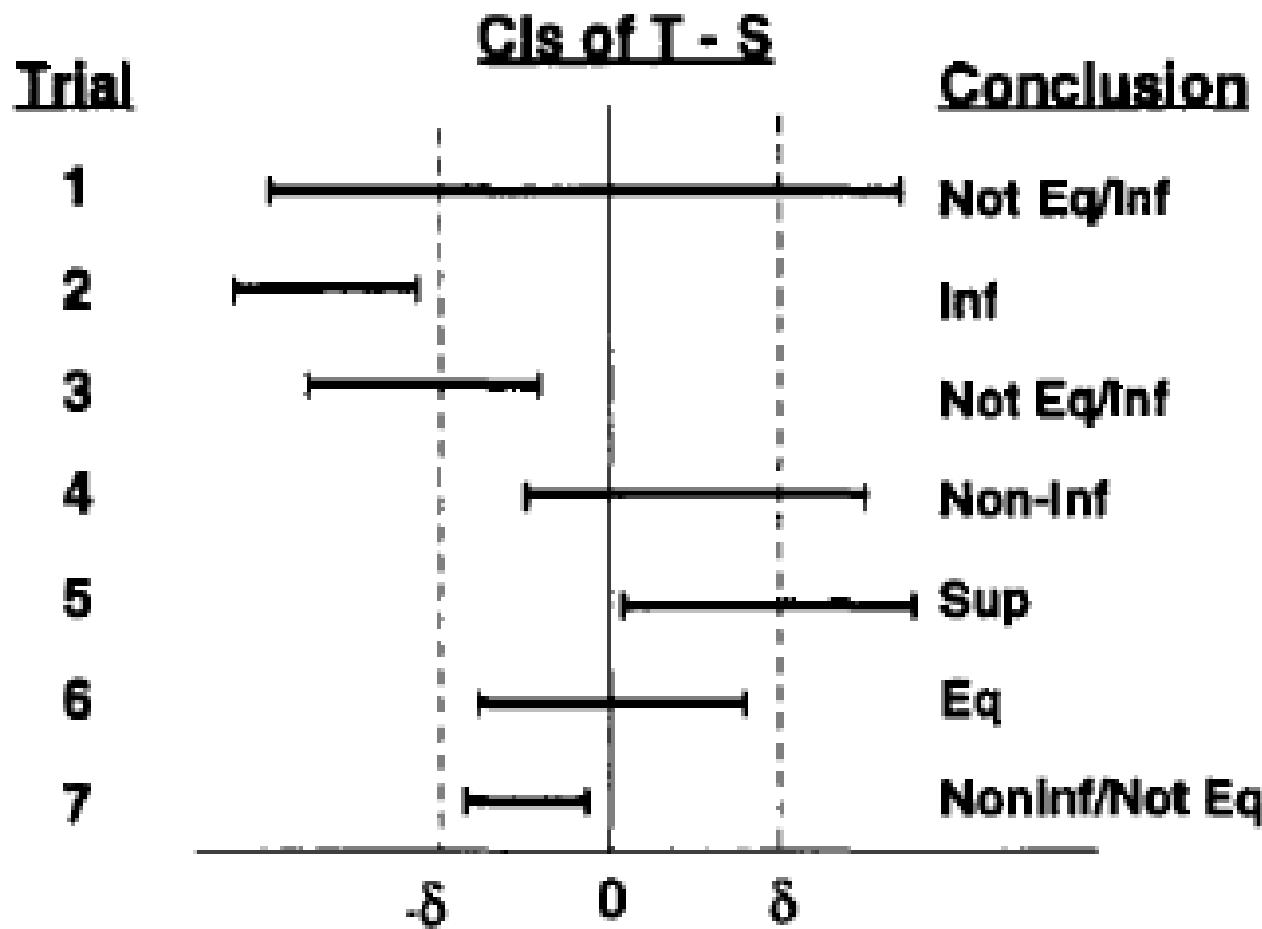


FIGURE 4. Using the CI approach to assess noninferiority, equivalence, and superiority.
Not Eq = Not equivalent; Inf = inferior; NonInf = Noninferior; Sup = superior; Eq = Equivalent.

Sample Statistics in a Parallel Design Trial

- Quantify primary endpoint
- Perform a two sample t-test with a pre-determined type I error level (alpha), usually 0.05
- T-test generates a test statistic
- Calculate a p-value for that test statistic based on degrees of freedom
- Assumptions
 - Data from both groups are normally distributed
 - Both groups are independent of one another, generated by randomization
 - Variances of the two groups are the same

If Assumption(s) violated...

- Cannot perform a parametric two sample t-test
- Must perform a Wilcoxon Rank Sum Test
 - When populations small (generally, $n < 30$)
 - When distribution of the population(s) are heavily skewed
- Rank Sum Test uses non-parametric approach
 - Population does not have to be normally distributed
 - Can still determine differences between two populations

Sample Statistics in a Crossover Design Trial

- Sometimes referred to as a 2x2 design (2 periods, 2 treatments)
- Each subject their own control (treatment effect estimated w/ higher precision)
 - However, that treatment effect must be reversible, so during washout phase between periods, biomarker can go back to “baseline” (carry over)
 - Assumptions: data is normally distributed, crossover differences are random
- Paired t-test
 - Ignores period effect and carry over effect; can determine difference b/w two means

Nonparametric Paired Analyses

- When data is skewed or not normally distributed
- Use Wilcoxon Signed Rank Test (different than Rank Sum Test for unpaired data)

Sample Statistics in a Factorial Design Trial

- When more than two groups, have more than two means to compare
 - Example: Phase I dose escalation with three or more dose levels
 - Cannot perform t-test
 - Must use an analysis of variance (ANOVA) to determine whether variances of each group are different
 - Broken down into variance within each group (MS_{WIT}) and between each group (MS_{BET})
 - F statistic is MS_{BET}/MS_{WIT}
 - If $MS_{BET} \sim MS_{WIT}$, then occurs by chance alone, and that there is NO difference between group means (Null hypothesis accepted)
 - If $MS_{BET} \gg MS_{WIT}$, then difference not by chance alone, suggesting that the group means ARE different (Null hypothesis rejected)
- Assumptions for ANOVA:
 - Biomarker measured from a random sample of population
 - Errors are normally and independently distributed
 - Measurement variance is constant
- When assumption(s) are violated, perform the nonparametric Kruskal-Wallis test

Global F test

- If p-value of ANOVA is significant (based on df, alpha), then need to perform multiple comparisons test
 - Ensures significance not due to inflated type I error
 - ANOVA just tells us that one or more groups are different than overall mean, but not WHICH group(s)
 - If have 3 groups, then have 3 comparisons
 - Group 1 vs 2, Group 1 vs 3, Group 2 vs 3
 - Adjust type I error (alpha) according to number of comparisons (c):
 - Generally $\sim 1-(1-\alpha)^c$
 - Bonferroni adjustment
 - Minimally Significant Difference
 - Tukey's method



No of comparisons (c)	Maximum probability of Type-I error
1	0.050
2	0.098
3	0.143
5	0.226
10	0.401

Sample Size Estimation

- When properly utilized, can:
 - Provide correct precision or power
 - Optimize subject recruitment
 - Optimize trial duration

Types of Biomarker Indications

1. Symptom / Sign Benefit

- Pain relief, angina relief, lowered blood pressure, improved cognition

2. Morbidity / Mortality

- Reverse disease
 - Treatment of pediatric leukemia, stroke
- Retard disease
 - Treatment of congestive heart failure, some cancers

Biomarkers for Dose Selection

- Biomarkers can be effectively used to improve dose/regimen selection
 - Dose ‘range’ for symptom/sign related indications
 - Exposure/response analyses in pivotal Phase 3 registration trials
 - Dose(s) for mortality/morbidity indications
- Value of biomarkers is enhanced for drugs with approved indications
 - New target populations, new dosing regiments, new routes of delivery

Initial Dosing Regimen Selection

- Dose selection based on biomarker effect is mostly (not definitively) due to:
 - Multiple putative mechanisms of action
 - Uncertain mechanism(s)
 - Highly variable response and/or effects
 - Small net effect on endpoints
 - Subjective clinical endpoints
 - Alzheimers' Disease Assessment Scale (ADAS), Hamilton Depression rating scale (HAM-D), Barthel Index (for functional ability after stroke), solid tumor measurements via imaging

Pivotal Phase 3 Trials

- Comparison by Indication Type

Factor	Symptom/Sign	Mortality/Morbidity
Trial Length	Shorter	Longer
Trial Size	Smaller	Larger
Net Effect	Larger	Smaller
Effect Variability	Smaller	Larger
Dose Range	Wider	Seldom > 1 dose

- Trials that test symptom/sign – type indicators of benefit might select a range of doses/regimens using biomarkers
- Trials that test mortality/morbidity – type indicators of benefit might **have** to select the dose/regimen using biomarkers

Value to Drug Development

- More informed dose/regimen selection
 - Could lead to increased trial success
- Quantitative analysis was critical
- Effective use of prior data for predictions
- Supports conduct of useful shorter-duration trials for future compounds in same mechanistic class that utilize same biomarker

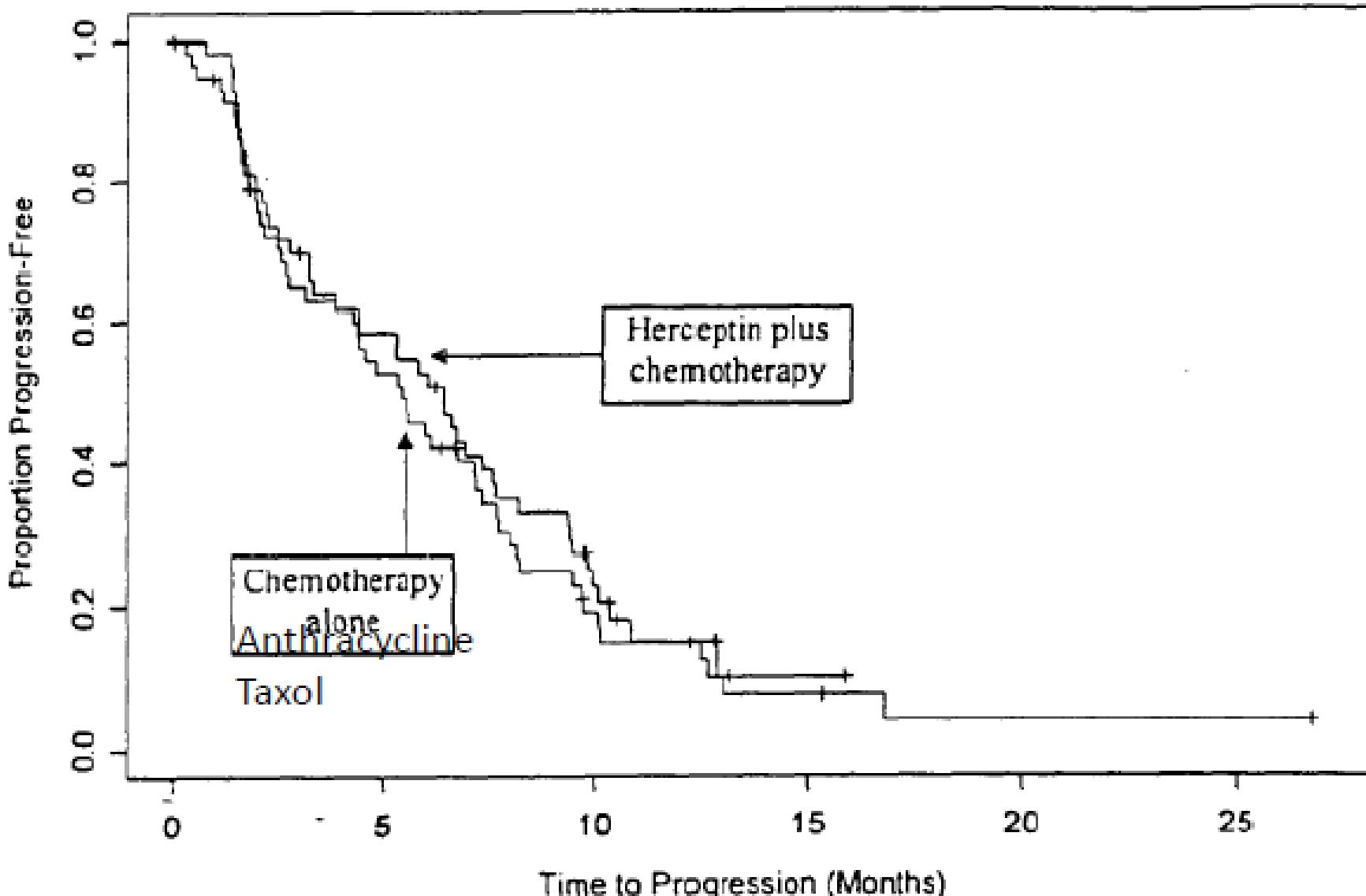
Role of Biomarkers for Target Population Selection

- Targeted therapies are designed to bind and alter function of a specific target
 - Typically a protein
 - Ex. Receptor, enzyme, transporter, kinase, phosphatase, etc
- Proteins are encoded by genes, which could vary by genotype within population
- Some people might not express gene/protein target
 - Therapy wouldn't work for them
- Use the target as a biomarker to select the “target population” for that therapy

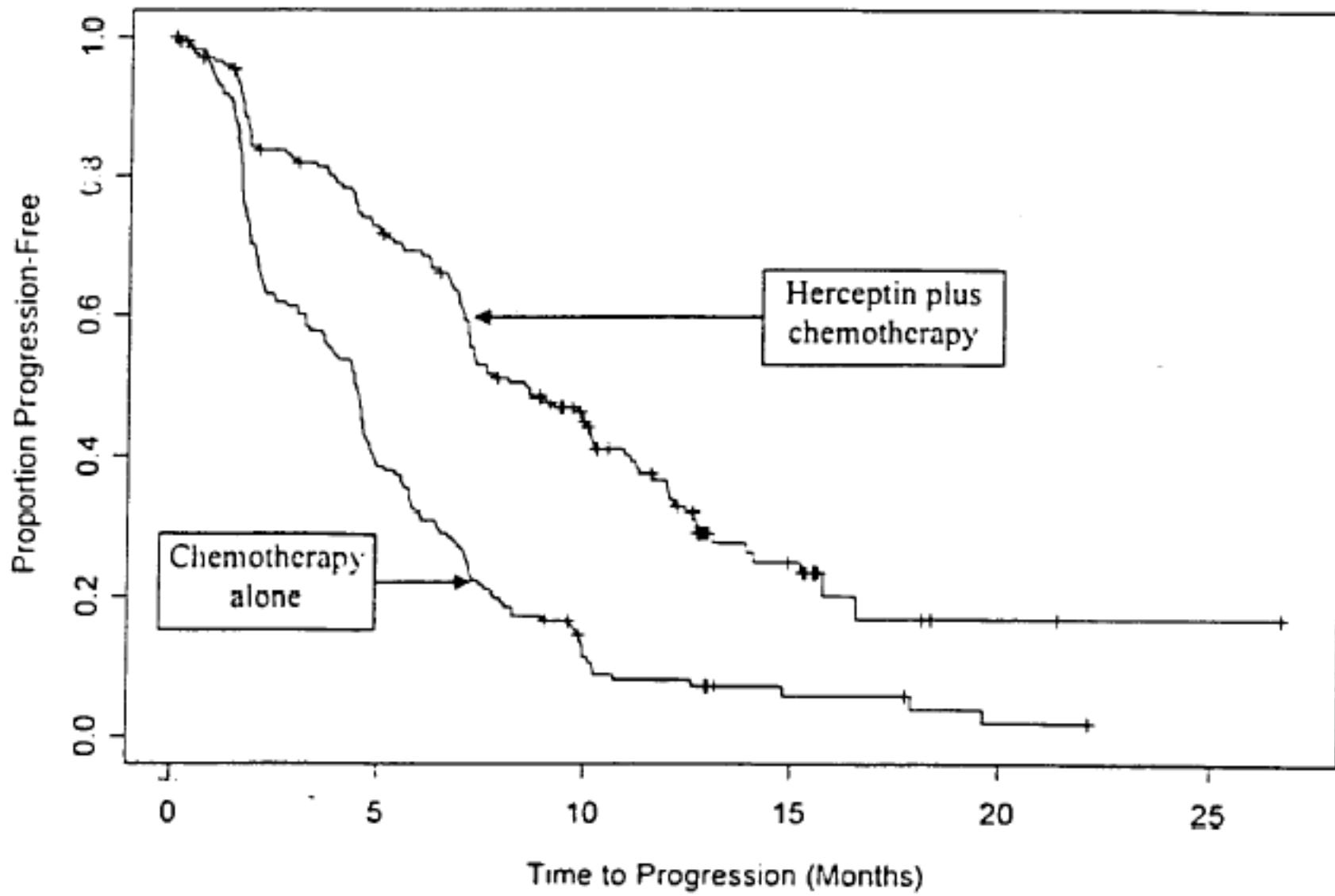
Herceptin (Trastuzumab)

- Recombinant DNA-derived humanized monoclonal antibody (mAb) that selectively binds to HER2
- HER2 overexpressed in some (not all) breast cancers
- Patients with tumors over-expressing HER2 protein would benefit most
 - HER2 protein score of 3+
- Patients with tumors that are HER2 negative, wouldn't benefit
- Identifying target patients can make the difference in drug's success or failure

Time to Progression Her 2+ Patients



Time to Progression Her 3+ Patients



AM Break

Exposure/Response Modeling I

Continuous Exposure vs Continuous Response Data

Types of Continuous Response Data

- Blood sugar
- Bacterial counts
- Tumor size
- Heart rate
- Blood pressure
- Forced expiratory volume in 1 second (FEV1)
 - For COPD diagnoses
- Prostate specific antigen (PSA)
- Blood counts
 - White blood cells (WBC)
 - Red blood cells (RBC)
 - Platelets
- Creatinine clearance
 - Kidney function
- Alanine aminotransferase
 - Liver function

*Suitable for use as clinical biomarkers for therapeutic efficacy or toxicity,
i.e. requires exposure/response modeling to assess optimal dose/exposure

Modeling Concepts

- Modeling is the process of estimating the parameters of a mathematical formulation that describes the relationship between two or more variables
 - Drug concentration vs time (cont vs cont; PK)
 - Dose vs PD response
 - Drug concentration (exposure) vs PD response
- Simulation is the use of a model to predict future outcomes
 - Altered dose regimen than what doses were observed
 - Different patient populations
 - Pediatrics, neo-nates, geriatrics, etc.
- Modeling and Simulation (M&S) of PK/PD has been proposed as a tool to improve efficiency of drug development

Regression Analysis

- Regression is a statistical technique used to explore relationships between quantities in a scientific system
- Linear regression is what we'll focus on here
- Uses a linear model
- Uses simple linear regression
 - Using ordinary least squares
 - Requires regression diagnostics to confirm accuracy of prediction and to satisfy certain assumptions that must be made
- All models have assumptions, therefore some error
“All models are wrong, but some are useful...”

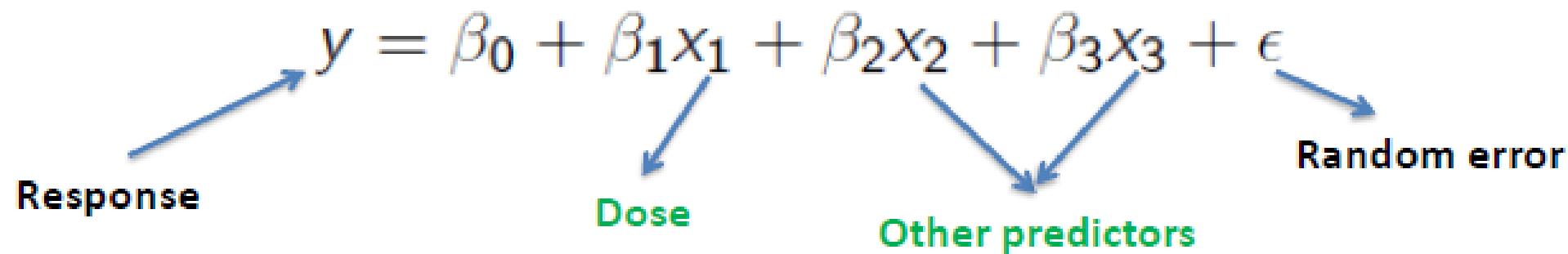
Understanding Dose-Response Relationships

- Knowing relationship between dose, exposure (drug concentration in blood following that dose) and clinical response (both effective and toxic) is important for determining safe and effective use of that drug
- Used to prepare dosage and administration instructions provided in the product label
- Used to choose a starting dose of a drug if know the shape and location of the population (or group) average dose-response curve
 - both for desirable and undesirable effects

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \epsilon$$

Linear Models

- A model is linear if the partial derivatives with respect to (wrt) any of the parameters are independent of the other parameters
- Ex:



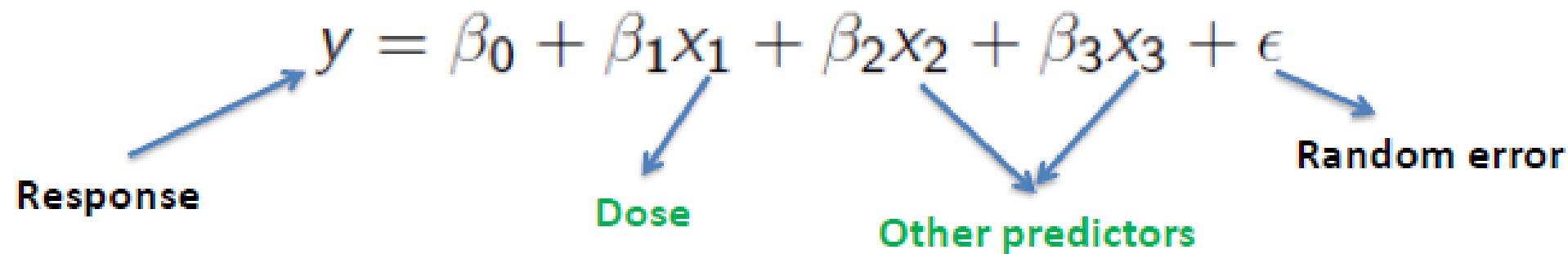
Partial Derivatives
Independent of other
parameters

$$\frac{\partial y}{\partial \beta_0} = 1$$

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \epsilon$$

Linear Models

- A model is linear if the partial derivatives with respect to (wrt) any of the parameters are independent of the other parameters
- Ex:



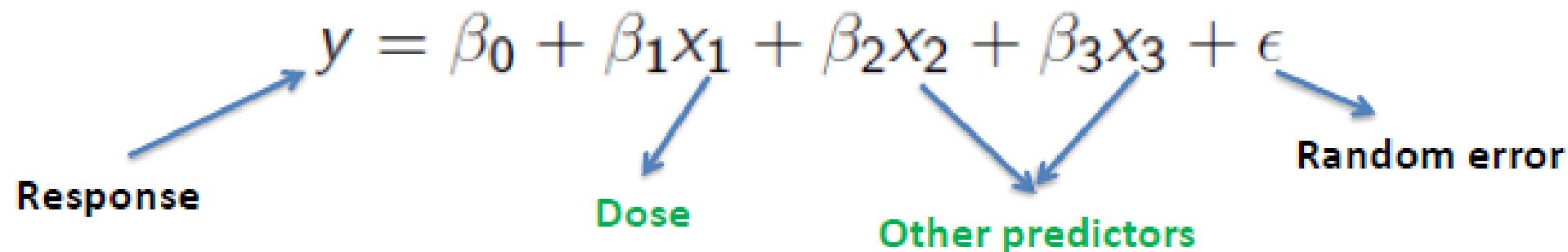
Partial Derivatives
Independent of other
parameters

$$\frac{\partial y}{\partial \beta_0} = 1 \quad \frac{\partial y}{\partial \beta_1} = x_1$$

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \epsilon$$

Linear Models

- A model is linear if the partial derivatives with respect to (wrt) any of the parameters are independent of the other parameters
- Ex:



Partial Derivatives
Independent of other
parameters

$$\frac{\partial y}{\partial \beta_0} = 1 \quad \frac{\partial y}{\partial \beta_1} = x_1 \quad \frac{\partial y}{\partial \beta_2} = x_2 \quad \dots \quad \frac{\partial y}{\partial \beta_k} = x_k$$

Linear Models

- Is this model linear?

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2^2 + \beta_3 x_2^3 + \epsilon$$

Partial Derivatives

Still independent of other parameters

$$\frac{\partial y}{\partial \beta_0} = 1 \quad \frac{\partial y}{\partial \beta_1} = x_1 \quad \frac{\partial y}{\partial \beta_2} = 2x_2 \quad \frac{\partial y}{\partial \beta_3} = 3x_2^2$$

- Yes, the model is still linear
- Even though there are square and cubic terms for predictor variables, model still linear wrt its parameters (betas) (aka regression coefficients)

Simple vs Multiple Linear Regression

- Simple linear regression: a response variable (dependent) is a function of only one predictor variable (independent)
 - Ex: $y = \beta_0 + \beta_1 x_1 + \epsilon$
- Multiple linear regression: a response variable (dependent) is a function of two or more predictor variables (independent)
 - Ex: $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \epsilon$

Simple Linear Regression

- Case Study: change in diastolic blood pressure from baseline (delta DBP or simply dDBP) with dose of a drug

$$dDBP = \beta_0 + \beta_1 Dose + \epsilon \quad (\text{generic})$$

$$dDBP_i = \beta_0 + \beta_1 Dose_i + \epsilon_i \quad (\text{specific for the } i^{\text{th}} \text{ subject})$$

- Let's let $y = dDBP$; $x = \text{Dose}$
- Collect trial data on 40 subjects $(x_1, y_1), (x_2, y_2), \dots, (x_{40}, y_{40})$

Simple Linear Regression

1. Read in dataset

```
. data<-read.csv("simpleLinReg_day2.csv", header=TRUE)
. str(data)
data.frame': 40 obs. of 3 variables:
$ ID : int 1 2 3 4 5 6 7 8 9 10 ...
$ dose: num 0 0 0 0 0 0 0 0 0 0 ...
$ resp: num -3.42 -11.82 -1.57 5.26 -1.27 ...
. head(data)
   ID   dose      resp
1  1     0 -3.418372
2  2     0 -11.822551
3  3     0 -1.566089
4  4     0  5.259403
5  5     0 -1.270960
6  6     0 -5.856467
.
```

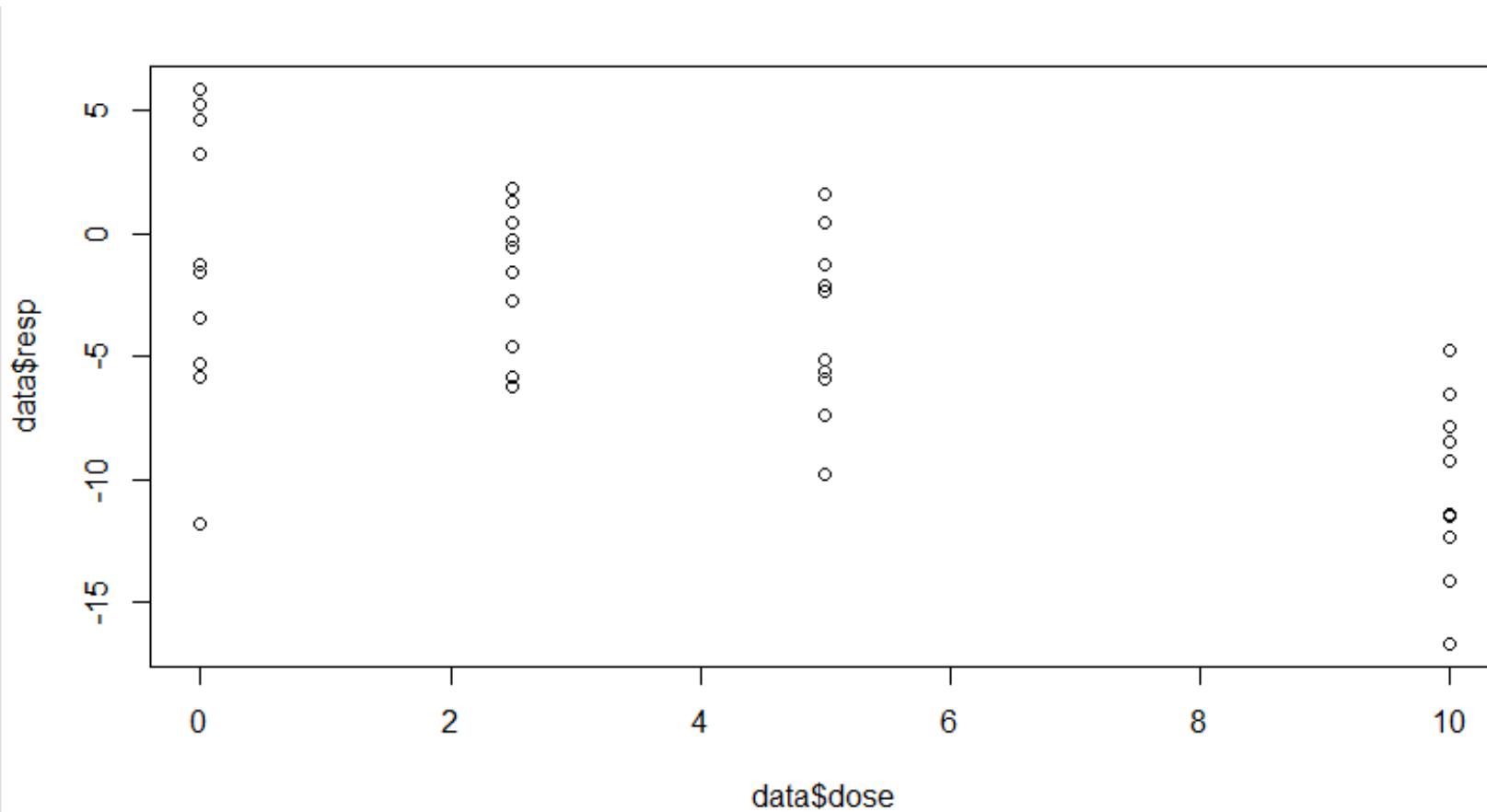
2. Get a sense of dataset structure

3. View “header” of dataset

Simple Linear Regression

4. Plot simple xy plot to view data

```
plot(data$dose, data$resp)
```

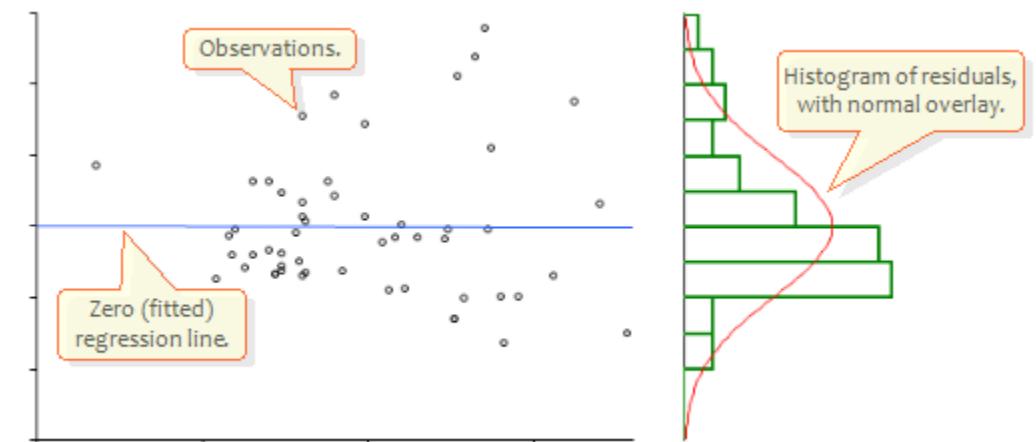
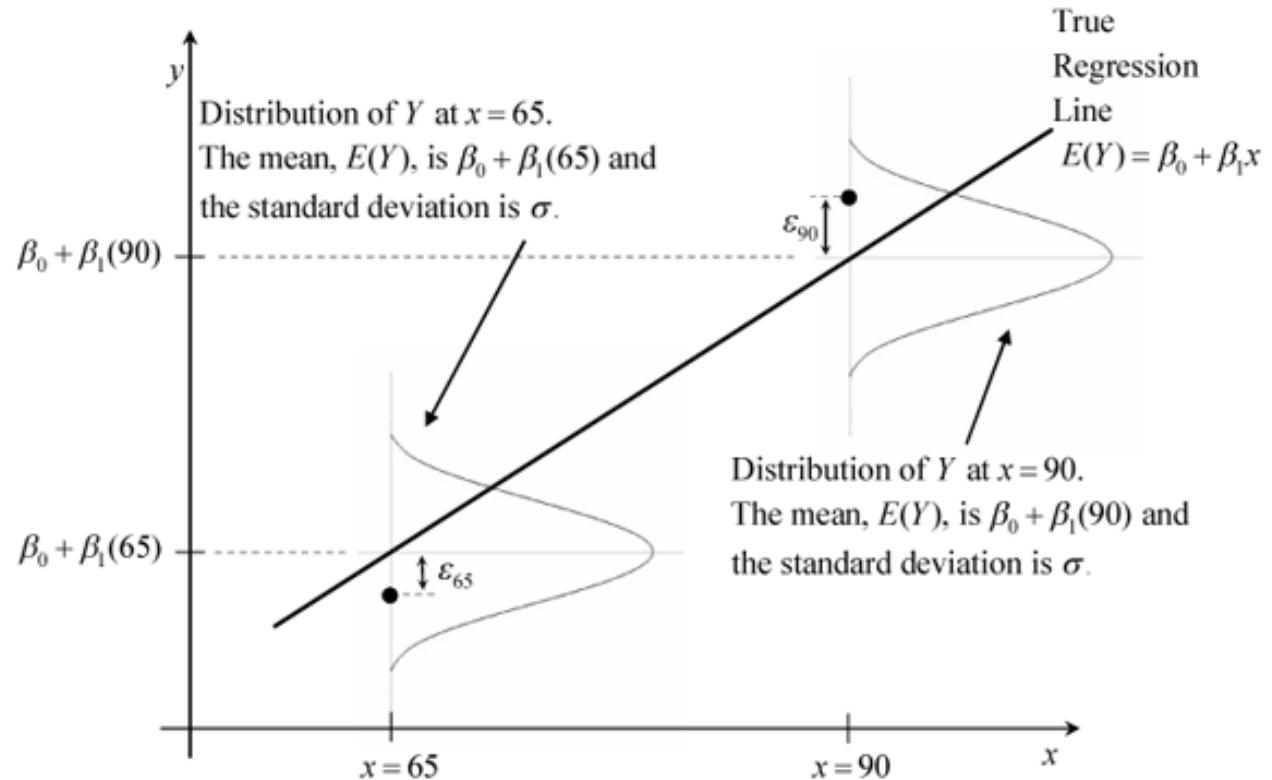


Simple Linear Regression

- Considerable inter-subject variability
- Purpose of the linear regression model is to provide a “best-fit” through the observed data
- Can then use that “best-fit” to **explain** behavior of system and/or **predict future** observations
- The subject-specific model:
 - Where ϵ_i is a residual error term
 - Normally distributed, centered on zero, with variance on either side (standard deviation (σ) squared)
 - $\epsilon_i \sim N(0, \sigma^2)$

Residual Error

- Observed data plotted and a model-predicted “best-fit” line is drawn
- The “residual” is the difference between observed and predicted
 - Denoted by ε

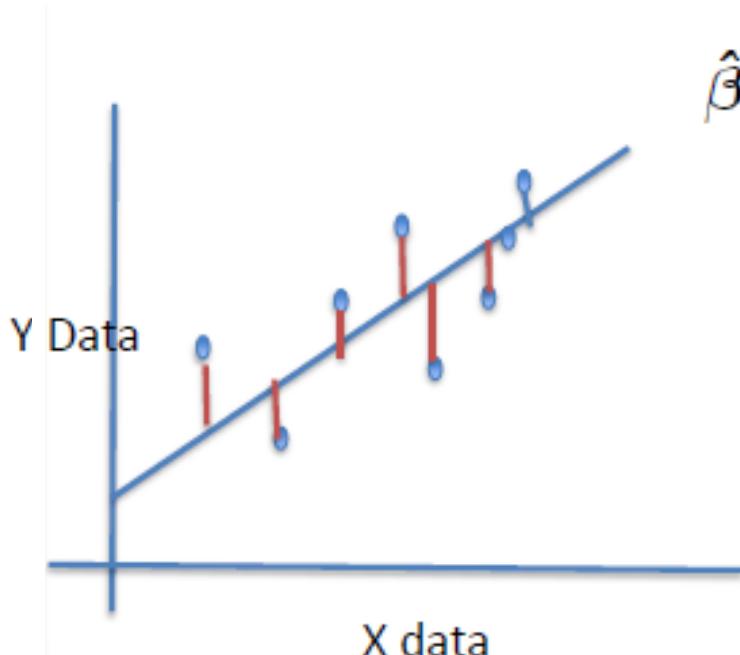


Simple Linear Regression

- The subject-specific model:
 - For subject i
$$y_i = \beta_0 + \beta_1 Dose_i + \epsilon_i$$
- The predicted model:
 - The “^” indicates model-predicted parameter
$$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 Dose + \hat{\epsilon}_i$$
- Uses ordinary least squares (OLS) to determine model regression
 - Each observation contributes equally to regression line
 - No weighting factors to mask influences of outliers
 - For each parameter (β_0 and β_1), a different value of “y” is tried (one iteration)
 - In that iteration using some value of “y”, a parameter value is tried and the residual determined
 - Tries multiple iterations

Simple Linear Regression

- The subject-specific model: $y_i = \beta_0 + \beta_1 Dose_i + \epsilon_i$
- The predicted model: $\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 Dose + \hat{\epsilon}_i$
- The value of β that produces the smallest residual ($obs - pred$; or $y - \hat{y}$) is the optimal value for that parameter.



$$\hat{\beta} : \min \sum_{i=1}^n e_i^2$$
$$\min \sum_{i=1}^n (y_i - \hat{y}_i)^2$$

i = iteration

Residual sum of squares
Or
Objective function
Or
Error sum of squares

Assumptions for OLS

- The independent variable (in this example, dose) is a *fixed* variable
 - There's no “error” in that variable; at least negligible error
- ϵ_i are normal random variables
 - $N(0, \sigma^2)$ is a constant variance, i.e. homogenous variance
- ϵ_i are uncorrelated from observation-to-observation

Linear Regression in R

$$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 Dose + \hat{\epsilon}_i$$

$$\hat{y}_i = (-0.02606) + (-0.9684)Dose + \hat{\epsilon}_i$$

- No special R package needed
- Results for parameter estimates
 - Since an “estimate”, there’s some uncertainty
 - Uncertainty measured by standard error (SE)
- Residuals presented as quartiles

```
> attach(data)
> myfit<-lm(resp~dose)
> summary(myfit)

Call:
lm(formula = resp ~ dose)

Residuals:
    Min      1Q  Median      3Q     Max 
-11.7965 -2.6303  0.0535  3.1062  6.3987 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept) -0.02606   1.00828  -0.026    0.98    
dose        -0.95837   0.17602  -5.445 3.29e-06 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 4.116 on 38 degrees of freedom
Multiple R-squared:  0.4382,    Adjusted R-squared:  0.4235 
F-statistic: 29.64 on 1 and 38 DF,  p-value: 3.287e-06
```

Understanding Output from R

$$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 Dose + \hat{\epsilon}_i$$

Coefficients:

	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	-0.02606	1.00828	-0.026	0.98
dose	-0.95837	0.17602	-5.445	3.29e-06 ***

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$$

$$\hat{\beta}_1 = \frac{\sum_{i=1}^n (x_i - \bar{x}) y_i}{\sum_{i=1}^n (x_i - \bar{x})^2} \quad \hat{\sigma}^2 = s^2 = \frac{\sum_{i=1}^n (y_i - \hat{y}_i)^2}{n - p}$$

Residual standard error: 4.116 on 38 degrees of freedom

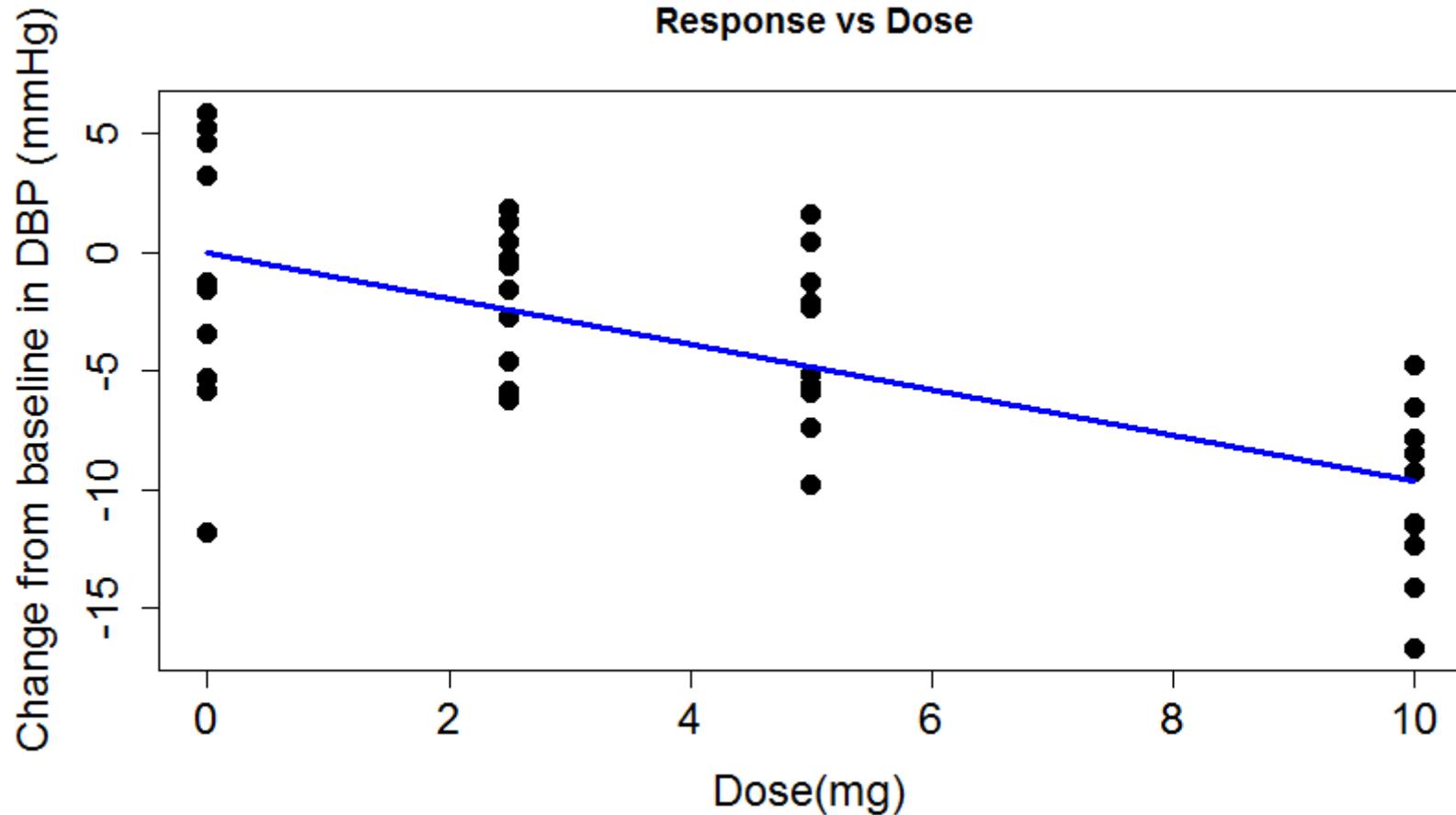
Multiple R-squared: 0.4382, Adjusted R-squared: 0.4235
F-statistic: 29.64 on 1 and 38 DF, p-value: 3.287e-06

Plot Linear Regression in R

```
resid(myfit)
fitted(myfit)

plot(dose,resp,cex = 1.5,cex.axis=1.5,cex.lab=1.5,ylab=
      "Change from baseline in DBP (mmHg)",xlab="Dose(mg)",pch = 16,
      main=" Response vs Dose")

lines(dose, fitted(myfit),lwd = 3.0,col="blue")
```



Interpretation of the Parameter Estimates

$$\hat{y}_i = (-0.02606) + (-0.9684)Dose + \hat{\epsilon}_i$$

- β_0 is the value of the mean response (dDBP) when the dose (independent variable, x) is zero
 - Aka the y-intercept
 - Where the best-fit line intersects with the y-axis (when x=0)
 - If there is NO placebo effect, the β_0 should be zero
 - Since $\beta_0 = -0.0206$, suggests there's **some** placebo effect
- β_1 is the change in the mean response (**mean** dDBP) per unit increase in the dose
 - Dose units are mg
 - For every 1 mg increase in drug dose, the mean dDBP will decrease by 0.968
 - Refer to the **mean** dDBP b/c the model is essentially an average of all observations

Inference on the Slope Parameter

- Is the slope of this regression line significantly different from zero?
- Null hypothesis states the line is NOT different from zero
 - A lack of difference in slope from zero suggests no drug effect on lowering dDBP
- Can use a t-test to determine if slope significantly different from zero
 - Original results already ran t-test and listed the p-value

```
Residuals:
    Min      1Q   Median      3Q      Max 
-11.7965 -2.6303  0.0535  3.1062  6.3987 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept) -0.02606    1.00828  -0.026    0.98    
dose        -0.95837    0.17602  -5.445 3.29e-06 *** 
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 4.116 on 38 degrees of freedom
Multiple R-squared:  0.4382,    Adjusted R-squared:  0.4235 
F-statistic: 29.64 on 1 and 38 DF,  p-value: 3.287e-06
```

Inference on the Slope Parameter

- Is the slope of this regression line significantly different from zero?
- Can use another method to determine statistical significance:
- Calculate the 95% confidence interval (CI) ($\alpha=0.05$):
 - $100(1-\alpha)\%$ CI for β_1 is:

$$\widehat{\beta}_1 \pm t_{\frac{\alpha}{2}, n-2} * SE(\widehat{\beta}_1)$$

- 95% CI does NOT contain zero !
- Slope is significantly different from Zero
 - There is a drug effect

```
> confint(myfit)
              2.5 %    97.5 %
(Intercept) -2.067212  2.0150932
dose        -1.314700 -0.6020347
```

Inference on the Slope Parameter

- Is the slope of this regression line significantly different from zero?
- Can also use an analysis of the variance (ANOVA) test
- Compares the variances of two or more groups
 - Variances are the measures of variability in model
 - Partition the total variability:

$$\sum_{i=1}^n (\hat{y}_i - \bar{y})^2 + \sum_{i=1}^n (y_i - \hat{y}_i)^2 = \sum_{i=1}^n (y_i - \bar{y})^2$$

Regression Sum of Squares (SS_{Reg})

Residual Sum of Squares (SSE)

Total Sum of Squares (SS_{Total})

Variability explained by the model:
(predicted – mean observed)

Unexplained variability:
Can't explain why that particular
observation different than predicted

Variability explained by the model:
(observed – mean observed)

ANOVA Approach to Regression

- Good model would have $SS_{Reg} \gg SSE$
 - Want the model to account for much more variability (SS_{Reg}) vs unexplained variability (SSE)
 - Conduct a Global F-test to test hypothesis:
 $H_0: \beta_1 = 0$ vs $H_A: \beta_1 \neq 0$
- $p = \# \text{ parameters (2)}$
 $n = \# \text{ observations (40)}$
 $n-p: \text{degrees of freedom}$

$$F = \frac{SS_{Reg}/P}{SSE/(n - p)} = \frac{MS_{Reg}}{MSE} \sim \text{Under } H_0, F_{p,n-p}$$

Reject H_0 if $F > F_{p,n-p, \alpha}$

ANOVA for Dose/Response Data in R

- Reject H_0 , concluding that there is a statistically significant linear trend between dose and response

```
-----  
> anova(myfit)  
Analysis of Variance Table  
  
Response: resp  
          Df Sum Sq Mean Sq F value    Pr(>F)  
dose      1 502.29  502.29  29.644 3.287e-06 ***  
Residuals 38 643.86   16.94  
---  
signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Interpretations of Global F test

- F-statistic given with linear model results
 - Global F-test determines a model's "goodness of fit", i.e. how well the model described the data
- Goodness of fit typically described by correlation coefficient of determination:
$$r^2 = \frac{SS_{Reg}}{SS_{Total}}$$
- Refers to the proportional of total variation in response (y) that is explained by the model
 - The higher, the better the model fit

```
call:
lm(formula = resp ~ dose)

Residuals:
    Min      1Q  Median      3Q     Max 
-11.7965 -2.6303  0.0535  3.1062  6.3987 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept) -0.02606   1.00828  -0.026    0.98    
dose        -0.95837   0.17602  -5.445 3.29e-06 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 4.116 on 38 degrees of freedom
Multiple R-squared:  0.4382,    Adjusted R-squared:  0.4235 
F-statistic: 29.64 on 1 and 38 DF,  p-value: 3.287e-06
```

Multiple Comparisons

- Because there were 4 different doses (0mg, 2.5mg, 5mg, and 10mg), we need to perform multiple comparisons
 - Compare each dose vs the other
- Use Tukey's method
 - Have to treat dose as a “**factor**”:
- 10mg significantly better than 0mg, 2.5mg, and 5mg
- No other dose was significantly greater than another
- Only the 10mg dose works

```
> TukeyHSD(aov(resp~as.factor(dose)))
Tukey multiple comparisons of means
 95% family-wise confidence level

Fit: aov(formula = resp ~ as.factor(dose))

$`as.factor(dose)`
      diff      lwr      upr   p adj
2.5-0 -0.7909054 -5.765606 4.183795 0.9732659
5-0   -2.7308223 -7.705523 2.243879 0.4606624
10-0  -9.2776725 -14.252373 -4.302972 0.0000800
5-2.5 -1.9399169 -6.914618 3.034784 0.7213428
10-2.5 -8.4867670 -13.461468 -3.512066 0.0002897
10-5   -6.5468502 -11.521551 -1.572149 0.0058586
```

Ultimate Goal

- What was the original research question?
 1. Did drug significantly lower dDBP?
- Yes, dose-related changes in response were found
 - The dose-response relationship followed a linear pattern
 - Slope of the line significantly different from zero, so a definite drug effect
- 2. Did we decide a dose that should be chosen for future studies?
- Yes, we showed that only 10mg of drug significantly lowered dDBP

Multiple Linear Regression

- Same principles as simple linear regression, just with multiple (2+) independent variables, each with their own estimable parameters (β_n)
- Example: drug clearance
 - Can be affected by body weight, age, and/or smoking status

$$y_i = \beta_0 + \beta_1 x_i + \beta_2 x_2 + \beta_3 x_3 + \epsilon_i$$

$$CL_i = \beta_0 + \beta_1 Weight_i + \beta_2 Age_i + \beta_3 Smoker_i + \epsilon_i$$

Multiple Linear Regression

- Case study: 26 patients with advanced carcinomas given 5-fluorouracil (5-FU) with methotrexate (MTX) once q2-3weeks
 - PK data collected on Day 1
 - 5-FU clearance values determined for each patient (noncompartmental analysis)
 - Also collected following covariate data from patients:
 - Age (yrs)
 - Sex (M/F)
 - BSA (m^2)
 - 5-FU dose (mg)
 - MTX (yes/no)

Multiple Linear Regression

- Case study: 26 patients with advanced carcinomas given 5-fluorouracil (5-FU) with methotrexate (MTX) once q2-3weeks

Goals:

1. Develop a useful model relating 5-FU clearance to patient demographics
2. Utilize model for individualized dosing regimens in future

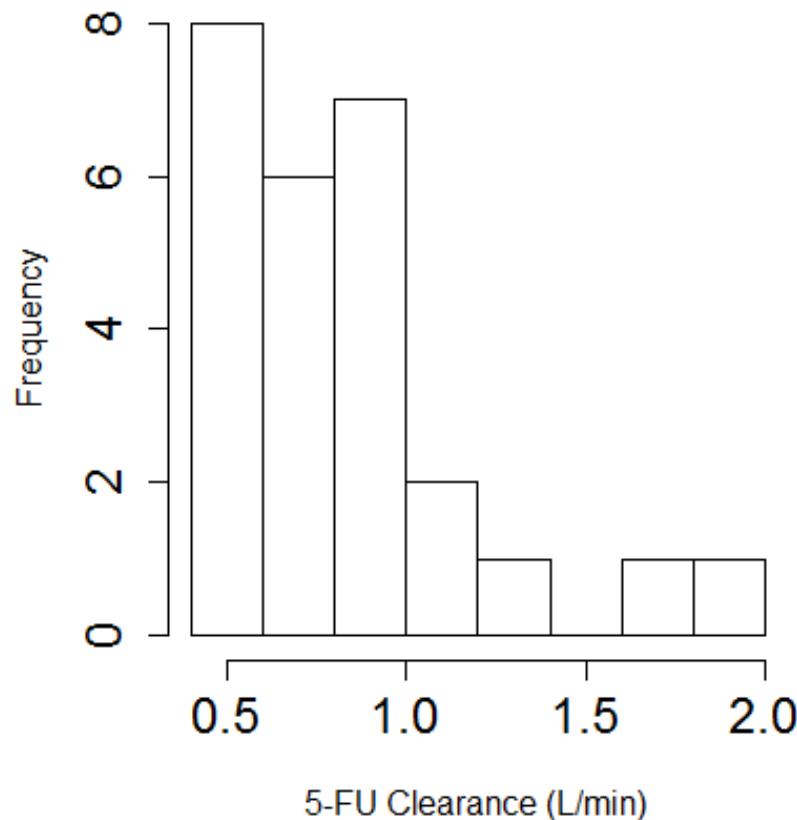
Multiple Linear Regression

```
> fudata<-read.csv("M1tp1LinReg_Day2.csv", header = TRUE)
> str(fudata)
'data.frame': 26 obs. of 7 variables:
 $ ID   : int  1 2 3 4 5 6 7 8 9 10 ...
 $ SEX  : int  1 1 1 0 1 1 1 0 0 0 ...
 $ AGE  : int  62 53 55 53 54 50 57 61 62 49 ...
 $ BSA   : num  1.65 1.63 2.14 2.14 1.91 1.66 1.6 2.05 1.94
 $ DOSE  : int  1500 750 1500 1800 1500 1500 1500 1600 850 1
 $ MTX   : int  1 0 1 1 1 1 1 0 1 ...
 $ FU_CL: num  0.58 0.56 0.47 0.85 0.73 0.71 0.61 0.86 1.36
> head(fudata)
  ID SEX AGE  BSA DOSE MTX FU_CL
1  1   1  62 1.65 1500   1  0.58
2  2   1  53 1.63  750   0  0.56
3  3   1  55 2.14 1500   1  0.47
4  4   0  53 2.14 1800   1  0.85
5  5   1  54 1.91 1500   1  0.73
6  6   1  50 1.66 1500   1  0.71
> |
```

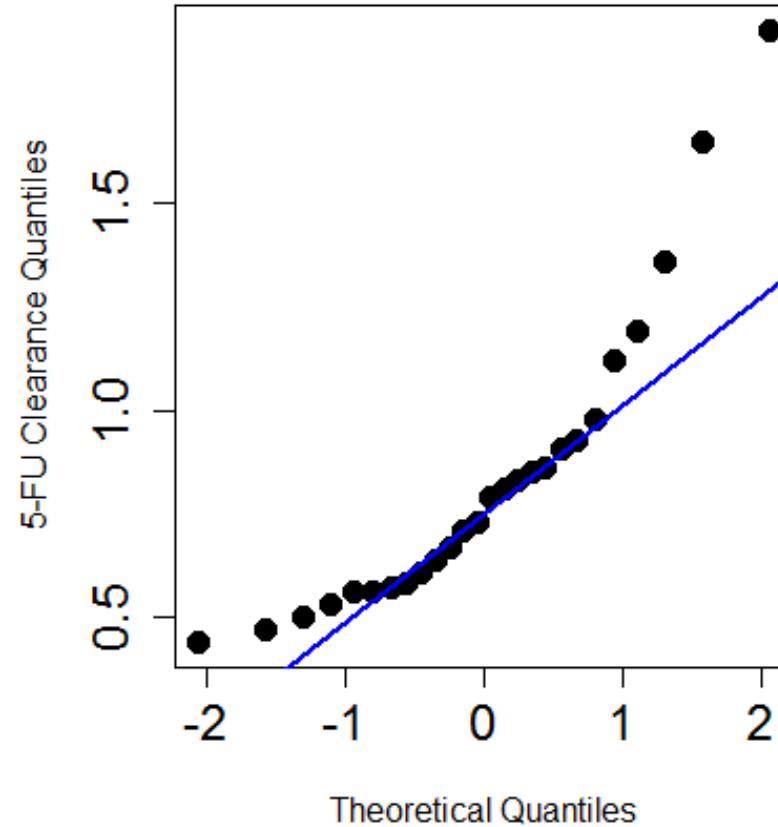
Exploratory Graphs

```
### exploratory plots ###
par(mfrow=c(1,2))
hist(FU_CL, breaks=10, xlab="5-FU clearance (L/min)", cex=1.5, cex.axis=1.5)
qqnorm(FU_CL, pch=16, cex=1.5, cex.axis=1.5, ylab = "5-FU clearance Quantiles")
qqline(FU_CL, col="blue", cex=1.5, lwd=2.0)
```

Histogram of FU_CL



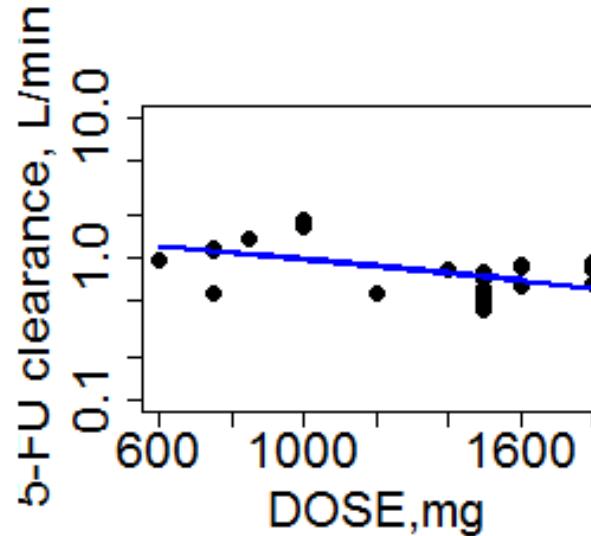
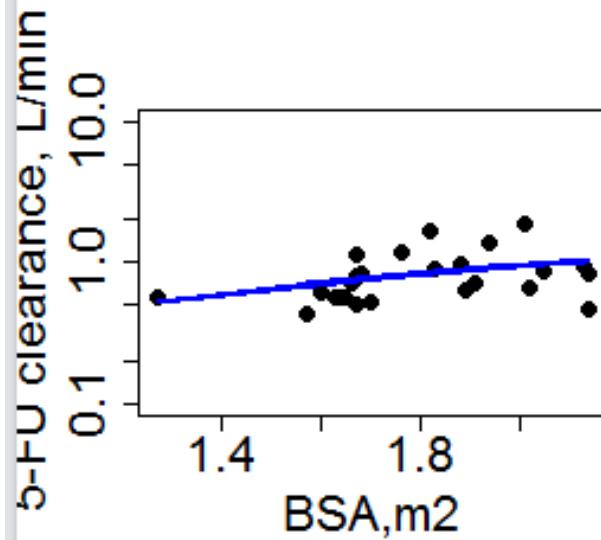
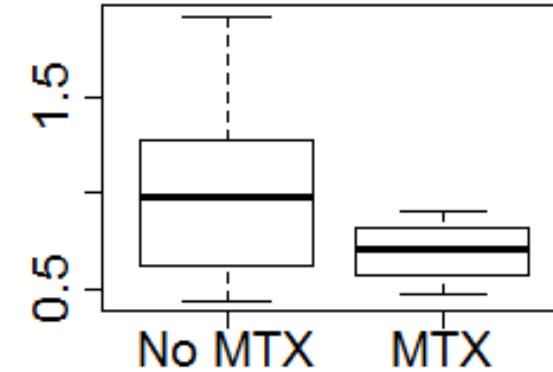
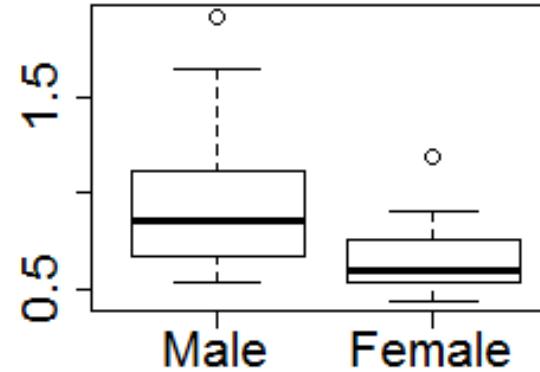
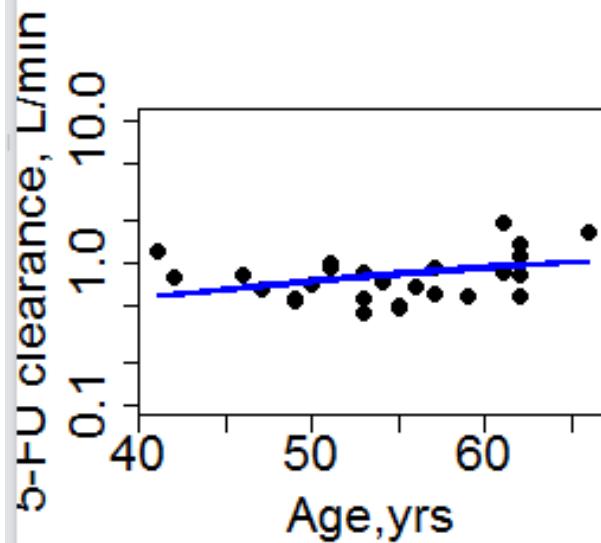
Normal Q-Q Plot



Exploratory Graphs of Covariates

```
### semi-log plots ###
par(mfrow=c(2,3))
plot(AGE,FU_CL,log="y",ylim=c(0.1,10),cex = 1.5,cex.axis=2.0,cex.lab=2.0,
     ylab="5-FU clearance, L/min",xlab="Age,yrs",pch = 16)
clf1<-lm(FU_CL~AGE)
lines(AGE, fitted(clf1),lwd = 2.0,col="blue")
boxplot(FU_CL~SEX,cex=1.5,cex.axis=2.0,cex.lab=2.0,names=c("Male","Female"))
boxplot(FU_CL~MTX,cex=1.5,cex.axis=2.0,cex.lab=2.0,names=c("No MTX","MTX"))
plot(BSA,FU_CL,log="y",ylim=c(0.1,10),cex = 1.5,cex.axis=2.0,cex.lab=2.0,
     ylab="5-FU clearance, L/min",xlab="BSA,m2",pch = 16)
clf2<-lm(FU_CL~BSA)
lines(BSA, fitted(clf2),lwd = 2.0,col="blue")
plot(DOSE,FU_CL,log="y", ylim=c(0.1,10),cex = 1.5,cex.axis=2.0,cex.lab=2.0,
     ylab="5-FU clearance, L/min",xlab="DOSE,mg",pch = 16)
clf3<-lm(FU_CL~DOSE)
lines(DOSE, fitted(clf3),lwd = 2.0,col="blue")
```

Exploratory Graphs of Covariates



- Slight increase in 5-FU clearance with Age
- Slight increase in 5-FU clearance with body size
- Slight decrease in 5-FU clearance with 5-FU dose
- Females have decreased 5-FU clearance vs Males
- MTX decreases 5-FU clearance vs No MTX

Is there any relationship between covariates?

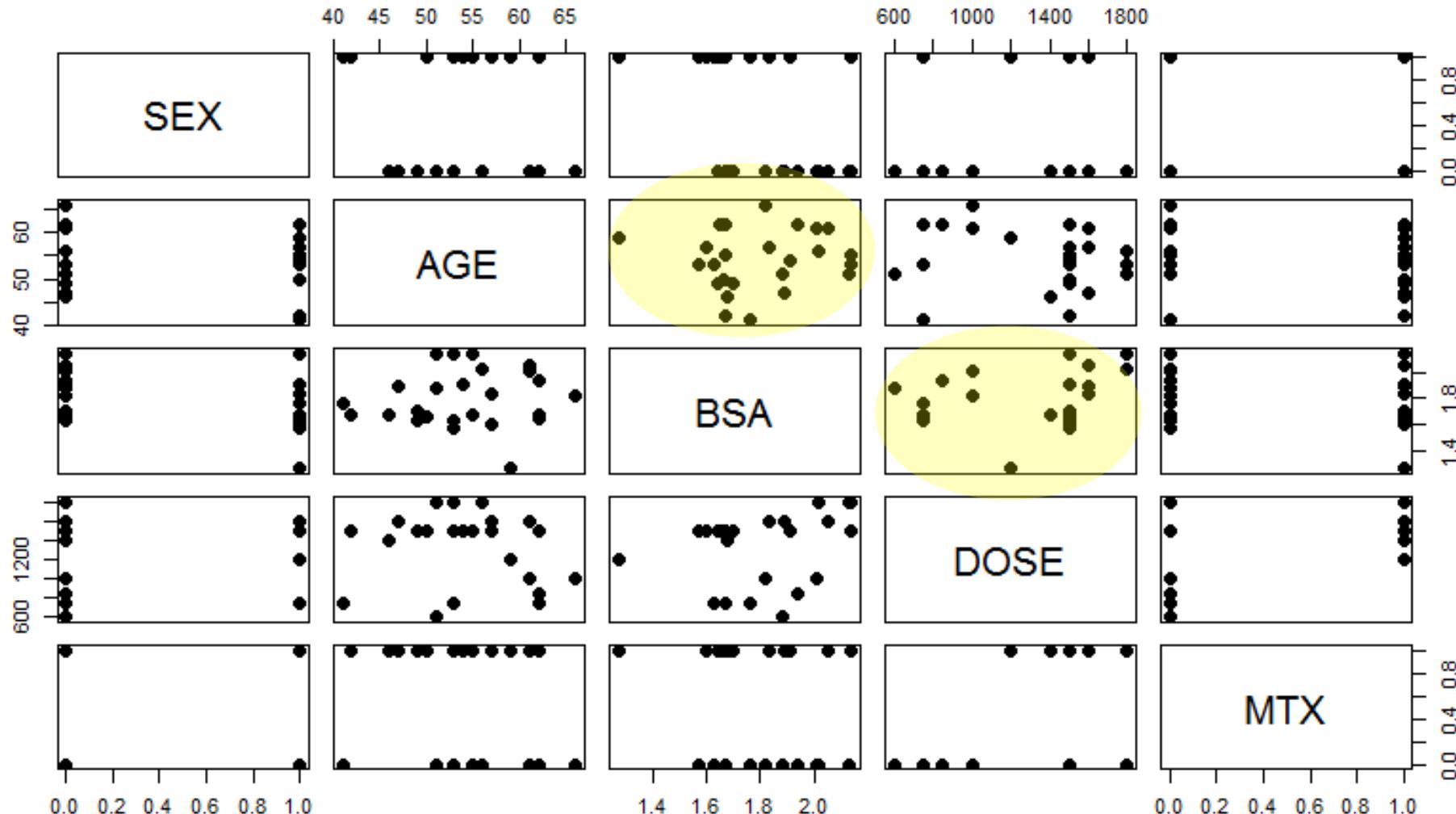
Relationship Among Covariates

Males = 0

Females = 1

```
### Relationship Among Covariates ###
pairs(fudata[,2:6], pch=16, cex=1.5)
```

"[,2:6]" refers to columns 2-6
If wanted specific rows, then
would've typed "[2:6,]"



Correlation Matrix

- Looking for $r > 0.4$
- Two pop out

1. BSA vs Sex

- Negative correlation
- Males = 0; Females = 1
- BSA decreases from “0” to “1”
- Meaning: males have larger BSA

2. MTX vs Dose

- Positive correlation
- The higher the 5-FU dose, the greater the odds of also taking MTX

```
> cor(fudata[,2:6])
```

	SEX	AGE	BSA	DOSE	MTX
SEX	1.00000000	-0.17692072	-0.42950109	0.03211843	0.1681750
AGE	-0.17692072	1.00000000	0.05757455	-0.11966129	-0.1559717
BSA	-0.42950109	0.05757455	1.00000000	0.22384062	-0.1452373
DOSE	0.03211843	-0.11966129	0.22384062	1.00000000	0.5560720
MTX	0.16817499	-0.15597171	-0.14523729	0.55607196	1.0000000

How did these r-values get calculated?

Pearson Correlation

$$\rho_{Y_1 Y_2} = \text{Cor}(Y_1 Y_2) = \frac{\text{Cov}(Y_1 Y_2)}{\sqrt{\text{Var}(Y_1)} * \sqrt{\text{Var}(Y_2)}}$$

- Pearson's product-moment correlation
- Calculates correlation between two variables (Y1 and Y2) using:
- The probability of two variables being correlated is equal to the covariance of the two variables divided by the product of the square root of each variable's variance

Pearson's product-moment correlation

```
data: MTX and DOSE
t = 3.2777, df = 24, p-value = 0.00318
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
0.2150383 0.7762279
sample estimates:
cor
0.556072
```

```
> cor(fudata[,2:6])
          SEX      AGE      BSA      DOSE      MTX
SEX  1.00000000 -0.17692072 -0.42950109  0.03211843  0.1681750
AGE -0.17692072  1.00000000  0.05757455 -0.11966129 -0.1559717
BSA -0.42950109  0.05757455  1.00000000  0.22384062 -0.1452373
DOSE  0.03211843 -0.11966129  0.22384062  1.00000000  0.5560720
MTX   0.16817499 -0.15597171 -0.14523729  0.55607196  1.0000000
```

Model Development

- Including a covariate (e.g. age, sex, BSA) should have physiological reasoning
- Univariate analysis (one variable at a time) helps identify how each individual covariate influences the response (in this case, CL)
 - Once that complete for each variable, begin model building using statistical approach
- A simpler model usually chosen over more complex models for ease of interpretation
- Log transform CL data to make linear regression easier to interpret the matrix design

$$\log(CL) = \beta_0 + \beta_1 Age + \beta_2 BSA + \beta_3 Dose + \beta_4 Sex + \beta_5 MTX + \epsilon$$

Matrix Notation

$$\log(CL) = \beta_0 + \beta_1 Age + \beta_2 BSA + \beta_3 Dose + \beta_4 Sex + \beta_5 MTX + \epsilon$$

$$\mathbf{y}_{nx1} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$$

Vector of response variables

Design matrix $nx(k+1)$

Vector of regression coefficients $(k+1)x1$

Vector of errors $nx1$

$$\begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix}_{nx1} = \begin{bmatrix} 1 & BSA_1 & Age_1 & DOSE_1 & 0 & 1 \\ 1 & BSA_2 & Age_2 & DOSE_2 & 1 & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ 1 & BSA_n & Age_n & DOSE_n & 0 & 0 \end{bmatrix}_{nx6} \begin{bmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \\ \beta_3 \\ \beta_4 \\ \beta_5 \end{bmatrix}_{6x1} + \begin{bmatrix} \epsilon_1 \\ \epsilon_2 \\ \vdots \\ \epsilon_n \end{bmatrix}_{nx1}$$

Ordinary Least Squares

- Find the set of predicted responses that are closes to observed responses by minimizing the residual sum of squares
 - Same concept as in simple linear regression
 - Multiple LR utilizes a matrix to accommodate 6 different variables influencing the response variable simultaneously

Univariate Analysis (Simple Linear Regression)

- 5-FU Clearance vs Sex:

```
### univariate analysis ####
### CL vs sex ##
plot(fudata$SEX, data$FU_CL)
attach(fudata)
LM_fu_sex<-lm(log(FU_CL) ~ factor(SEX))
summary(LM_fu_sex)
resid(LM_fu_sex)
fitted(LM_fu_sex)
plot(SEX,FU_CL,cex = 1.5,cex.axis=1.5,cex.lab=1.5,ylab=
      "5-FU clearance (L/min)",xlab="SEX",pch = 16)
lines(SEX, fitted(LM_fu_sex),lwd = 3.0,col="blue")
```

Residuals:

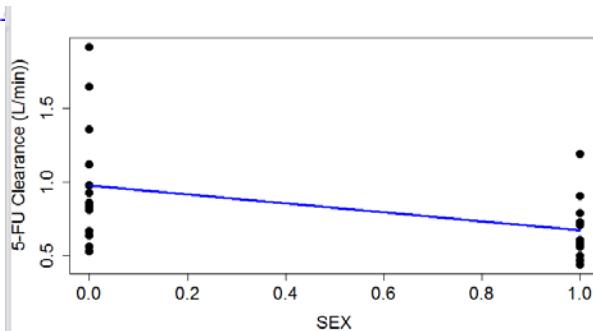
Min	1Q	Median	3Q	Max
-0.54268	-0.22620	-0.06448	0.18315	0.74452

Coefficients:

	Estimate	Std. Error	t value	Pr(> t)		
(Intercept)	-0.09219	0.09157	-1.007	0.3241		
factor(SEX)1	-0.34642	0.13479	-2.570	0.0168 *		

Signif. codes:	0 ‘***’	0.001 ‘**’	0.01 ‘*’	0.05 ‘.’	0.1 ‘ ’	1

Residual standard error: 0.3426 on 24 degrees of freedom
 Multiple R-squared: 0.2158, Adjusted R-squared: 0.1831
 F-statistic: 6.605 on 1 and 24 DF, p-value: 0.0168



- 5-FU Clearance vs Dose:

```
### CL vs Dose ##
plot(fudata$DOSE, data$FU_CL)
attach(fudata)
LM_fu_dose<-lm(log(FU_CL) ~ DOSE)
summary(LM_fu_dose)
resid(LM_fu_dose)
fitted(LM_fu_dose)
plot(DOSE,FU_CL,cex = 1.5,cex.axis=1.5,cex.lab=1.5,ylab=
      "5-FU clearance (L/min)",xlab="Dose (mg)", pch = 16)
lines(DOSE, fitted(LM_fu_dose),lwd = 3.0,col="blue")
```

Residuals:

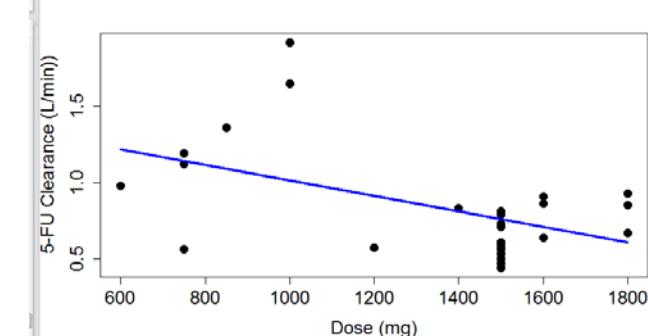
Min	1Q	Median	3Q	Max
-0.62905	-0.24121	0.03961	0.20335	0.72945

Coefficients:

	Estimate	Std. Error	t value	Pr(> t)		
(Intercept)	0.4283010	0.2642287	1.621	0.1181		
DOSE	-0.0005054	0.0001899	-2.661	0.0137 *		

Signif. codes:	0 ‘***’	0.001 ‘**’	0.01 ‘*’	0.05 ‘.’	0.1 ‘ ’	1

Residual standard error: 0.34 on 24 degrees of freedom
 Multiple R-squared: 0.2278, Adjusted R-squared: 0.1957
 F-statistic: 7.081 on 1 and 24 DF, p-value: 0.01367



Univariate Analysis (Simple Linear Regression)

- 5-FU Clearance vs Age:

```
### CL vs Age ##
plot(fudata$AGE, data$FU_CL)
attach(fudata)
LM_fu_age<-lm(log(FU_CL) ~ AGE)
summary(LM_fu_age)
resid(LM_fu_age)
fitted(LM_fu_age)
plot(AGE,logFUCL,cex = 1.5,cex.axis=1.5,cex.lab=1.5, ylab=
      "5-FU Clearance (L/min)",xlab="Age (yrs)", pch = 16)
lines(AGE, fitted(LM_fu_age),lwd = 3.0,col="blue")
```

Residuals:

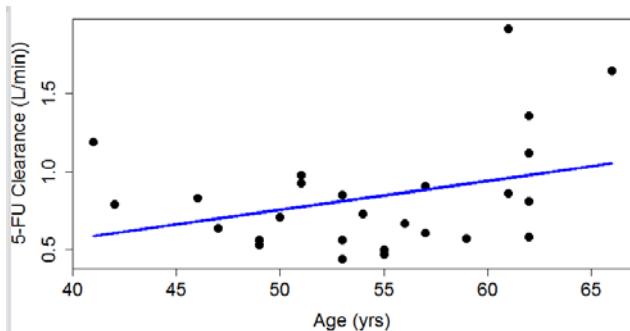
Min	1Q	Median	3Q	Max
-0.54661	-0.29316	-0.03813	0.22944	0.79789

Coefficients:

	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	-1.12772	0.62660	-1.800	0.0845 .
AGE	0.01610	0.01144	1.407	0.1722

Signif. codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.3719 on 24 degrees of freedom
Multiple R-squared: 0.0762, Adjusted R-squared: 0.03771
F-statistic: 1.98 on 1 and 24 DF, p-value: 0.1722



- 5-FU Clearance vs BSA:

```
### CL vs BSA ##
plot(fudata$BSA, data$FU_CL)
attach(fudata)
LM_fu_bsa<-lm(log(FU_CL) ~ BSA)
summary(LM_fu_bsa)
resid(LM_fu_bsa)
fitted(LM_fu_bsa)
plot(BSA,FU_CL,cex = 1.5,cex.axis=1.5,cex.lab=1.5, ylab=
      "5-FU Clearance (L/min)",xlab="BSA (m**2)", pch = 16)
lines(BSA, fitted(LM_fu_bsa),lwd = 3.0,col="blue")
```

Residuals:

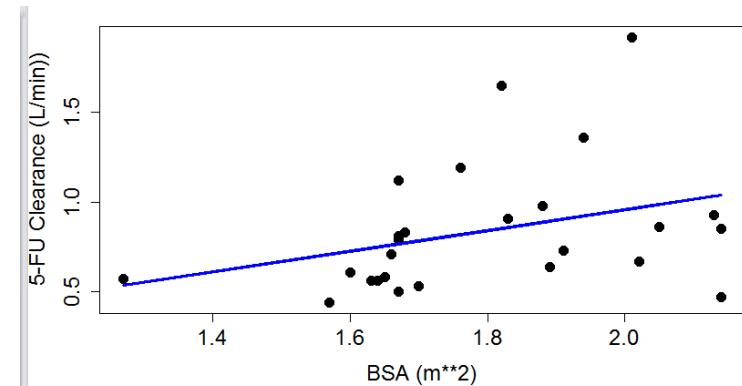
Min	1Q	Median	3Q	Max
-0.72872	-0.22721	-0.05293	0.13733	0.76304

Coefficients:

	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	-1.4159	0.6195	-2.286	0.0314 *
BSA	0.6494	0.3434	1.891	0.0707 .

Signif. codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.361 on 24 degrees of freedom
Multiple R-squared: 0.1297, Adjusted R-squared: 0.09342
F-statistic: 3.576 on 1 and 24 DF, p-value: 0.07074



Univariate Analysis (Simple Linear Regression)

- 5-FU Clearance vs MTX:

```
### CL vs MTX ##
plot(fudata$MTX, data$FU_CL)
attach(fudata)
LM_fu_mtx<-lm(log(FU_CL) ~ factor(MTX))
summary(LM_fu_mtx)
resid(LM_fu_mtx)
fitted(LM_fu_mtx)
plot(MTX,FU_CL,cex = 1.5,cex.axis=1.5,cex.lab=1.5, ylab=
  "5-FU clearance (L/min)",xlab="MTX", pch = 16)
lines(MTX, fitted(LM_fu_mtx),lwd = 3.0,col="blue")
```

Residuals:

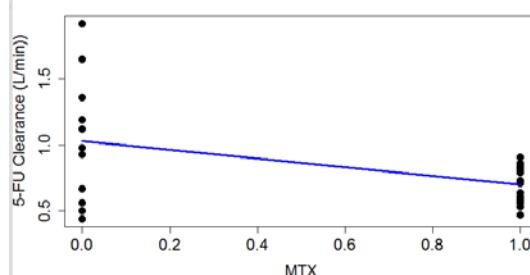
Min	1Q	Median	3Q	Max
-0.7447	-0.1944	0.0473	0.2125	0.7286

Coefficients:

	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	-0.0763	0.1067	-0.715	0.4813
factor(MTX)1	-0.3047	0.1404	-2.169	0.0402 *

Signif. codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.3538 on 24 degrees of freedom
Multiple R-squared: 0.164, Adjusted R-squared: 0.1291
F-statistic: 4.707 on 1 and 24 DF, p-value: 0.04017



Summary of Univariate Analyses

Variable	Intercept	SE	Slope	SE	R ²	P-value
Age	-1.128	0.627	0.016	0.011	0.038	0.172
BSA	-1.416	0.619	0.649	0.343	0.093	0.071
Dose	0.428	0.264	-0.000505	0.00018	0.196	0.014
Sex	-0.092	0.092	-0.346	0.135	0.183	0.017
MTX	-0.076	0.107	-0.305	0.140	0.129	0.040

*Can scale the continuous covariates to get more informative estimates

Scaling Covariates

- Scale continuous covariates
- Re-run using same codes, just scaled columns

```
### Scaling of continuous covariates ###
  ### Dose, scaled by 1000 #####
fudata$DOSE1<-DOSE/1000
fudata$AGE1<-AGE/50
fudata$BSA1<-BSA/1.7
```

```
> head(fudata)
  ID SEX AGE BSA DOSE MTX FU_CL DOSE1 AGE1      BSA1
1  1   1  62 1.65 1500   1  0.58  1.50 1.24 0.9705882
2  2   1  53 1.63  750   0  0.56  0.75 1.06 0.9588235
3  3   1  55 2.14 1500   1  0.47  1.50 1.10 1.2588235
4  4   0  53 2.14 1800   1  0.85  1.80 1.06 1.2588235
5  5   1  54 1.91 1500   1  0.73  1.50 1.08 1.1235294
6  6   1  50 1.66 1500   1  0.71  1.50 1.00 0.9764706
```

```
### scaled univariate analysis #####
### CL vs Dose ##
attach(fudata)
detach(fudata)
LM_fu_dose2<-lm(log(FU_CL) ~ fudata$DOSE1)
summary(LM_fu_dose2)

### CL vs Age ##
LM_fu_age2<-lm(log(FU_CL) ~ AGE1)
summary(LM_fu_age2)

### CL vs BSA ##
LM_fu_bsa2<-lm(log(FU_CL) ~ BSA1)
summary(LM_fu_bsa2)
```

Summary of *Scaled* Univariate Analyses

Variable	Intercept	SE	Slope	SE	R ²	P-value
Age	-1.128	0.627	0.805	0.572	0.038	0.172
BSA	-1.416	0.619	1.104	0.584	0.093	0.071
Dose	0.428	0.264	-0.505	0.189	0.196	0.014
Sex	-0.092	0.092	-0.346	0.135	0.183	0.017
MTX	-0.076	0.107	-0.305	0.140	0.129	0.040

Best Covariate Model

- Model selection performed to obtain the “best” set of predictors
 - Physiological/mechanistic reasoning
 - Statistical significance
- Automated sequential variable selection procedure
 - Backward elimination
 - Forward selection
 - Stepwise regression

Start with the “full” model, i.e. the model with all available covariates

Full Model

```
#####
##### Full model #####
full_lm<-lm(fudata$logFUCL ~ fudata$DOSE1 + fudata$AGE1 + fudata$BSA1 + factor(fudata$SEX) +
  factor(fudata$MTX))
summary(full_lm)
```

Residuals:

Min	1Q	Median	3Q	Max
-0.54192	-0.15973	0.01905	0.15806	0.42323

Coefficients:

	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	-1.01755	0.78140	-1.302	0.2076
fudata\$DOSE1	-0.57225	0.20767	-2.756	0.0122 *
fudata\$AGE1	0.42409	0.45073	0.941	0.3580
fudata\$BSA1	1.10078	0.55313	1.990	0.0604 .
factor(fudata\$SEX)1	-0.20127	0.12722	-1.582	0.1293
factor(fudata\$MTX)1	0.01191	0.14516	0.082	0.9354

Signif. codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.2855 on 20 degrees of freedom
Multiple R-squared: 0.5464, Adjusted R-squared: 0.433
F-statistic: 4.818 on 5 and 20 DF, p-value: 0.00472

$$H_0: \beta_1 = \beta_2 = \beta_3 = \beta_4 = \beta_5 = \beta_6 = 0$$

✓ $H_A: \text{At least one } \beta \neq 0$

Next step: remove variable with highest p-value
(least significant predictor)
= MTX

Backward Elimination

```
red_1m1<-lm(logFUCL ~ fudata$DOSE1 + fudata$AGE1 + fudata$BSA1 + factor(fudata$SEX))  
summary(red_1m1)
```

```
Residuals:  
    Min      1Q  Median      3Q     Max  
-0.53590 -0.16631  0.02033  0.15213  0.42206  
  
Coefficients:  
            Estimate Std. Error t value Pr(>|t|)  
(Intercept) -1.0076    0.7534  -1.337  0.19542  
fudata$DOSE1 -0.5621    0.1625  -3.459  0.00235 **  
fudata$AGE1   0.4211    0.4385   0.960  0.34781  
fudata$BSA1   1.0878    0.5172   2.103  0.04770 *  
factor(fudata$SEX)1 -0.2010    0.1241  -1.619  0.12031  
---  
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1  
  
Residual standard error: 0.2786 on 21 degrees of freedom  
Multiple R-squared:  0.5462,   Adjusted R-squared:  0.4598  
F-statistic: 6.32 on 4 and 21 DF,  p-value: 0.001679
```

Dose even more significant now
BSA also now a significant predictor

Age is now least significant predictor, so
remove in next iteration of model

Backward Elimination

```
red_lm2<-lm(logFUCL ~ fudata$DOSE1 + fudata$BSA1 + factor(fudata$SEX))
summary(red_lm2)
```

Residuals:

Min	1Q	Median	3Q	Max
-0.51925	-0.19700	0.04354	0.12039	0.46258

Coefficients:

	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	-0.5225	0.5580	-0.936	0.3593
fudata\$DOSE1	-0.5799	0.1611	-3.599	0.0016 **
fudata\$BSA1	1.0929	0.5163	2.117	0.0458 *
factor(fudata\$SEX)1	-0.2191	0.1225	-1.789	0.0874 .

Signif. codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.2781 on 22 degrees of freedom
Multiple R-squared: 0.5263, Adjusted R-squared: 0.4617
F-statistic: 8.148 on 3 and 22 DF, p-value: 0.0007835

Dose even more significant now
BSA slightly more significant

Sex is now least significant predictor, so
remove in next iteration of model

Backward Elimination

```
red_lm3<- lm(fudata$logFUCL ~ fudata$DOSE1 + fudata$BSA1)  
summary(red_lm3)
```

```
Residuals:  
    Min      1Q  Median      3Q     Max  
-0.71547 -0.11646  0.04638  0.15246  0.51302
```

```
Coefficients:  
            Estimate Std. Error t value Pr(>|t|)  
(Intercept) -1.0036    0.5118  -1.961  0.06209 .  
fudata$DOSE1 -0.6219    0.1669  -3.727  0.00111 **  
fudata$BSA1   1.5070    0.4831   3.119  0.00482 **  
---  
signif. codes:  0 '****' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
Residual standard error: 0.2911 on 23 degrees of freedom  
Multiple R-squared:  0.4574, Adjusted R-squared:  0.4102  
F-statistic: 9.694 on 2 and 23 DF, p-value: 0.0008842
```

Dose even more significant now
BSA slightly less significant

This appears to be an optimal version of the model

$$\log(CL) = -1.004 + 1.51 * \left(\frac{BSA}{1.7} \right) - 0.622 * \left(\frac{Dose}{1000} \right)$$

Interpretation of Model Parameters

$$\log(CL) = -1.004 + 1.51 * \left(\frac{BSA}{1.7} \right) - 0.622 * \left(\frac{Dose}{1000} \right)$$

- For a subject with a **fixed dose**, β_1 (intercept for BSA, 1.51) is interpreted as the mean change in clearance for a unit change in BSA
 - Has units of L*m²/min
- For a subject with a **fixed BSA**, β_2 (intercept for Dose, -0.622) is interpreted as the mean change in clearance for a unit (mg) change in dose
 - Has units of L*mg/min

Interpretation of Categorical Covariates

- Sex was not a statistically-significant covariate in this model
 - But let's pretend it is for now...
- The sex variable was listed as a binary (0 or 1) categorical covariate
 - Males = 0
 - Females = 1
- If sex was a significant covariate, how would we interpret the results?

Interpretation of Categorical Covariates

- 5-FU Clearance vs Sex:

```
### univariate analysis ####
### CL vs sex ##
plot(fudata$SEX, data$FU_CL)
attach(fudata)
LM_fu_sex<-lm(log(FU_CL) ~ factor(SEX))
summary(LM_fu_sex)
resid(LM_fu_sex)
fitted(LM_fu_sex)
plot(SEX,FU_CL,cex = 1.5,cex.axis=1.5,cex.lab=1.5,ylab=
  "5-FU clearance (L/min)",xlab="SEX",pch = 16)
lines(SEX, fitted(LM_fu_sex),lwd = 3.0,col="blue")
```

Residuals:

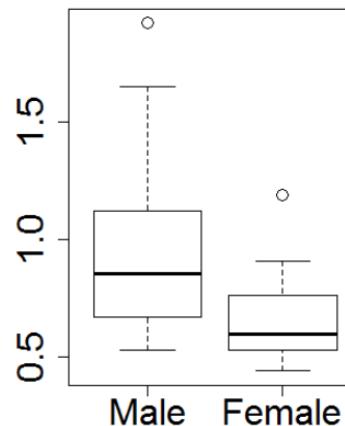
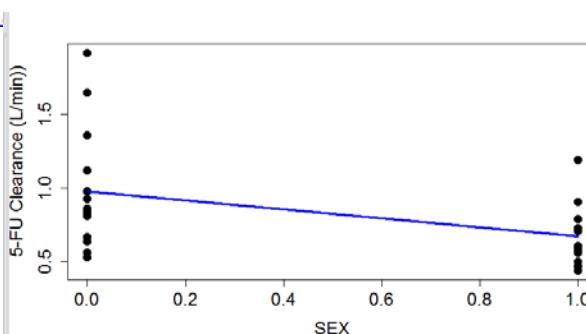
Min	1Q	Median	3Q	Max
-0.54268	-0.22620	-0.06448	0.18315	0.74452

Coefficients:

	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	-0.09219	0.09157	-1.007	0.3241
factor(SEX)1	-0.34642	0.13479	-2.570	0.0168 *

Signif. codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.3426 on 24 degrees of freedom
Multiple R-squared: 0.2158, Adjusted R-squared: 0.1831
F-statistic: 6.605 on 1 and 24 DF, p-value: 0.0168



- Results from the univariate model
- By itself, was a significant predictor
 - When variable used as a ‘factor’
 - Incorrect to treat binary 0 or 1 as a continuous covariate
 - To make model treat sex as categorical, must use the ‘factor’

Interpretation of Categorical Covariates

Simulate 40 random selections of 0 or 1

Bind that data in a column in the dDBP dataset

Run a multiple LR model with dose and sex (as factor)

```
### bivariate analysis ####
### dose/response with dDBP ##
### add 40 random 0 or 1 (males=0, females=1)
Sex<-rbinom(40,1,0.50)
data$SEX<-Sex
head(data)
Sex1<-factor(data$SEX)
attach(data)
drlm<-lm(resp ~ dose + Sex1)
summary(drlm)
```

Linear fit is given as:

$$\hat{y} = 1.6093 - 0.9855(Dose) - 2.6376(Sex)$$

For males (Sex = 0)

$$\hat{y} = 1.6093 - 0.9855(Dose)$$

For females (Sex = 1)

$$\hat{y} = 1.6093 - 2.6376 - 0.9855(Dose)$$

$$\hat{y} = -1.0283 - 0.9855(Dose)$$

```
Call:
lm(formula = resp ~ dose + Sex1)
```

Residuals:

Min	1Q	Median	3Q	Max
-13.4318	-2.3663	0.1861	3.1139	5.6907

Coefficients:

	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	1.6093	1.2452	1.292	0.2042
dose	-0.9855	0.1693	-5.822	1.09e-06 ***
Sex11	-2.6376	1.2660	-2.083	0.0442 *

Signif. codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 3.946 on 37 degrees of freedom

Multiple R-squared: 0.4972, Adjusted R-squared: 0.47

F-statistic: 18.3 on 2 and 37 DF, p-value: 2.989e-06

Interpretation of Categorical Covariates

**Categorical covariates shift the intercept, not the slope

Linear fit is given as:

$$\hat{y} = 1.6093 - 0.9855(Dose) - 2.6376(Sex)$$

For males (Sex = 0)

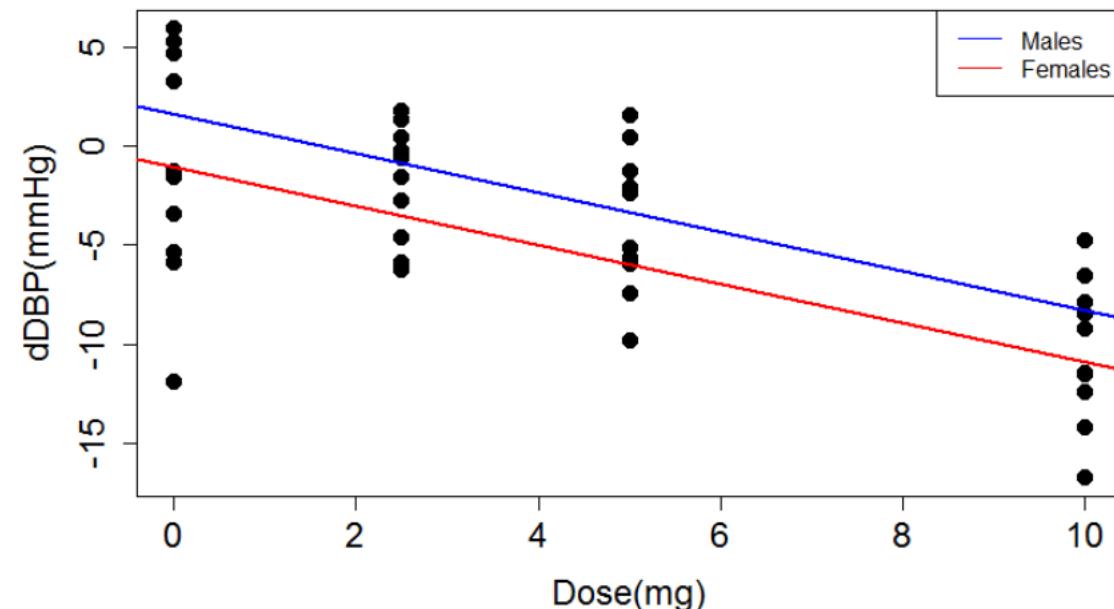
$$\hat{y} = 1.6093 - 0.9855(Dose)$$

For females (Sex = 1)

$$\hat{y} = 1.6093 - 2.6376 - 0.9855(Dose)$$

$$\hat{y} = -1.0283 - 0.9855(Dose)$$

```
plot(dose, resp, cex = 1.5, cex.axis=1.5, cex.lab=1.5, ylab=
      "dDBP(mmHg)", xlab="Dose(mg)", pch = 16)
abline(1.6093, -0.9855, col="blue", lwd=2.5)
(1.6093)+(-2.6376)
abline(-1.0283, -0.9855, col="red", lwd=2.5)
legend("topright", lwd=c(1,1), col=c("blue", "red"),
       legend = c("Males", "Females"))
```



Is this a meaningful model?

$$\log(CL) = -1.004 + 1.51 * \left(\frac{BSA}{1.7} \right) - 0.622 * \left(\frac{Dose}{1000} \right)$$

- BSA is a significant predictor of 5-FU clearance
 - As BSA increased, so too did 5-FU clearance
 - Makes mechanistic sense, as subjects with higher BSA are larger individuals
 - Larger subjects have larger drug-clearing organs with enzymes
 - In this case, dihydropyrimidine dehydrogenase (DPYD)
 - Yes, makes sense
 - Could've included Sex ($p=0.08$), but since $p>0.05$ and known to be related to BSA, doesn't add anything beyond BSA covariate
- Dose is a significant predictor of 5-FU clearance
 - As Dose increased, 5-FU clearance *decreased*
 - Could be saturating a metabolic/clearance pathway
 - As increase dose, aka amount of 5-FU molecules present, there's a "back-up": decreased CL
 - Yes, makes sense

Other Ways to Select Best Model

- Using model selection criterion
 - Akaike Information Criterion (AIC)
 - $AIC = -2 * \text{log-likelihood} + 2p$
 - P = number of predictor variables
 - Bayes Information Criterion (BIC)
 - $BIC = -2 * \text{log-likelihood} + 2p \log n$
 - n = number of observations or responses
 - The lower the AIC or BIC, the better the model fit

Collinearity and Ill-Conditioning

- When a large # of predictors present, problems of multicollinearity arise
 - Predictor variables may be correlated among themselves
 - Leads to unstable parameter estimates and high standard errors (SE) around those parameter estimates
 - AKA ill-conditioning

How to Detect Collinearity

- Correlation matrix of the covariates
 - High positive or negative correlation coefficients are indicative of collinearity
- Condition number (K)
 - Ratio of largest to smallest eigen value of the $X'X$ matrix
 - $K < 10^4$: no collinearity
 - $10^4 < K < 10^6$: moderate collinearity
 - $K > 10^6$: severe collinearity

How to Remove Collinearity

- Scaling, standardizing, and/or centering of covariates (predictor variables)
 - We did this for the continuous covariates (Age, BSA, Dose)
- Transform the collinear variables to another variable, then use that as the predictor variable
 - EX: instead of using both height and weight (which are often correlated), calculate BSA and use that

Let's Check for Collinearity in the 5-FU Data

- Using *unscaled* covariates, $K > 10^6$, suggesting severe collinearity

```
> mm<-model.matrix(~+AGE+BSA+DOSE+SEX+MTX)
> kappa(t(mm) %*% mm)
[1] 318720098
```

- Using *scaled* covariates, $K < 10^3$ (927), suggesting no collinearity

```
> mm1<-model.matrix(~ + fudata$AGE1 + fudata$BSA1 + fudata$DOSE1 + fudata$SEX + fudata$MTX)
> kappa(t(mm1) %*% mm1)
[1] 927.4529
```

- Using *scaled* covariates from final model, $K < 10^3$ (536), suggesting no collinearity

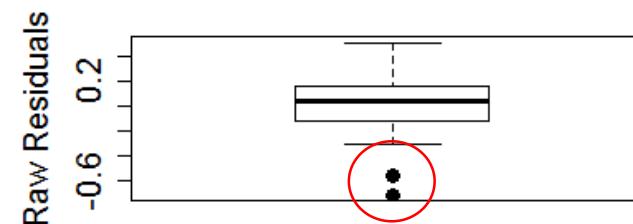
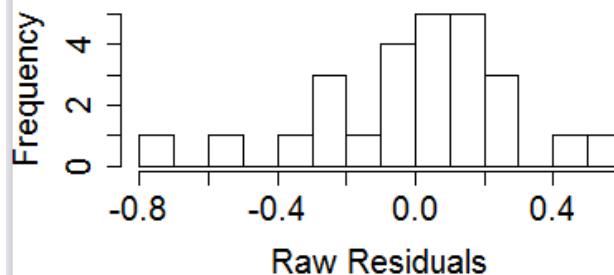
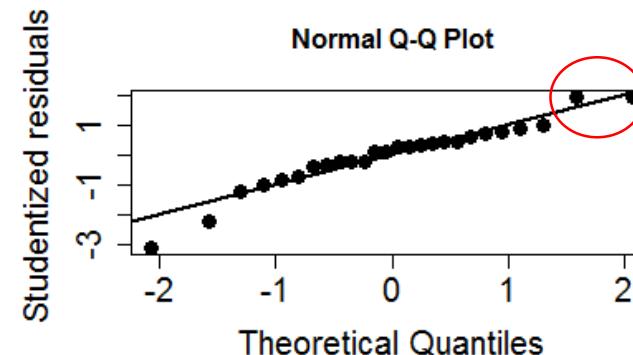
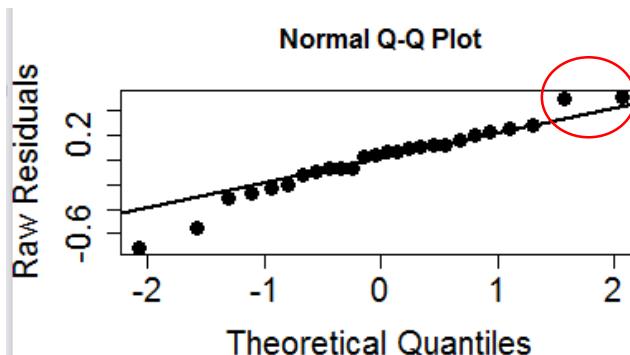
```
> mm2<-model.matrix(~ + fudata$BSA1 + fudata$DOSE1)
> kappa(t(mm2) %*% mm2)
[1] 536.2346
```

Regression Diagnostics

1. Linear regression have some assumptions that are made when building the model
 - Normal distribution of random error
 - Variance of parameter estimates is constant (homoscedasticity)
2. To verify these assumptions are satisfied, must run diagnostics
 - If don't, then model may be telling you biased information

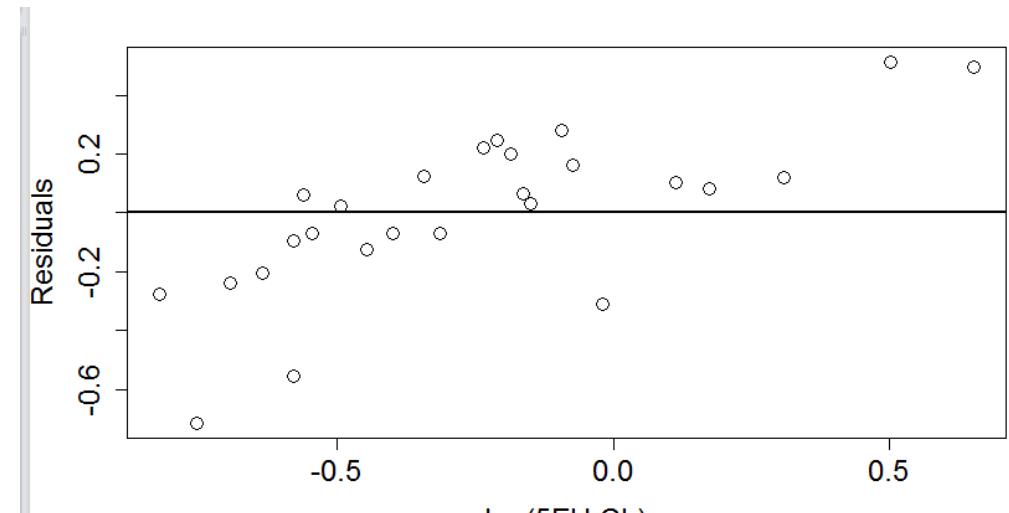
1. Assessing Normality of Error Distribution

```
#### Check Regression Diagnostics, assess normality of error assumptions--  
par(mfrow=c(2,2))  
qqnorm(red_lm3$res,ylab="Raw Residuals",cex=1.5,cex.axis=1.5,cex.lab=1.5,pch=16)  
qqline(red_lm3$res,lwd=2.0)  
qqnorm(rstudent(red_lm3),ylab=" Studentized residuals",pch=16,cex=1.5,cex.axis=1.5,  
abline(0,1,lwd=2.0)  
  
### Draw histogram of residuals, should be around zero  
hist(red_lm3$res,10,cex=1.5,cex.axis=1.5,cex.lab=1.5,xlab="Raw Residuals",main="")  
### Boxplot of residuals, should be around zero-  
boxplot(red_lm3$res,cex=1.5,cex.axis=1.5,cex.lab=1.5,pch=16,ylab="Raw Residuals")
```



2. Assessing Constant Variance Assumption

- Plot residuals vs log response
- Residuals are (for the most part) randomly scattered around zero
- Assumption satisfied
- If assumption violated, residuals would appear in funnel shape



```
#### Check constant variance assumption---  
### Plot residual on y-axis, predicted Log(CL) on x-axis, residuals should be zero--  
par(mfrow=c(1,2))  
plot(fudata$logFUCL, red_lm3$res, cex=1.5,cex.axis=1.5,cex.lab=1.5,xlab="log(5FU CL)",  
      ylab="Residuals")  
abline(0,0,lwd=2.0)
```

Influence Diagnostics

- Look for statistical outliers
 - In the X-direction: use process called Leverage
 - In the Y-direction: use standardized residuals
- To identify an influential observation, use Cook's Distance

Leverage

- Observations from some predictor variables (covariates) with high leverage can influence parameter estimates

$$\hat{y} = X\hat{\beta}$$

$$\hat{y} = X(X'X)^{-1}X'y$$

$$\hat{y} = Hy$$

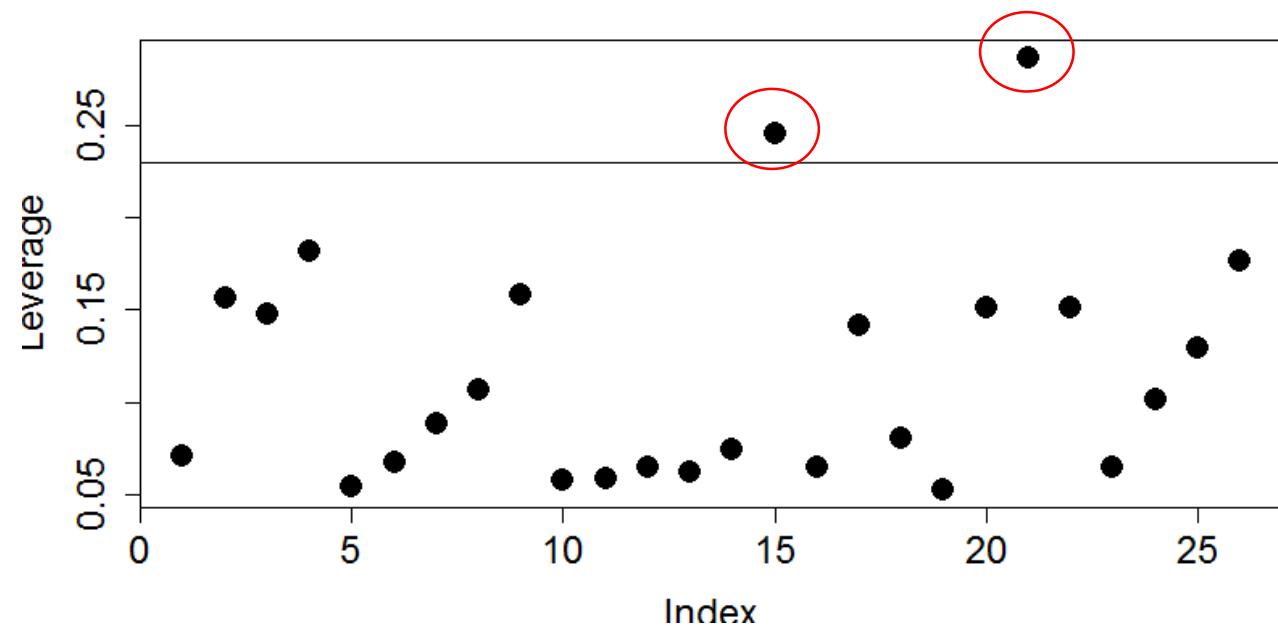
Where H is a Hat matrix

- Rule of thumb:

Leverage $> 2*p/n$ are influential

- In our case, $n=26$, $p=3$
- Leverage > 0.23 are influential

```
#### calculate leverage (influence in x-direction) with "hat function"--  
x<-model.matrix(red_lm3)  
lev<-hat(x)  
  
### plot leverage--  
par(mfrow=c(1,1))  
plot(lev,ylab="Leverage",cex=2.0,cex.lab=1.5,cex.axis = 1.5,pch = 16)  
names(lev)<-ID  
lev[lev>0.23]  
### Rule of thumb, if lev>2*p/n, then influential---
```

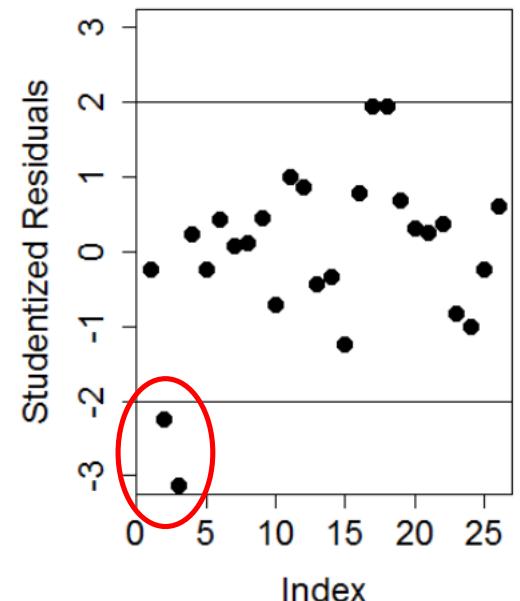
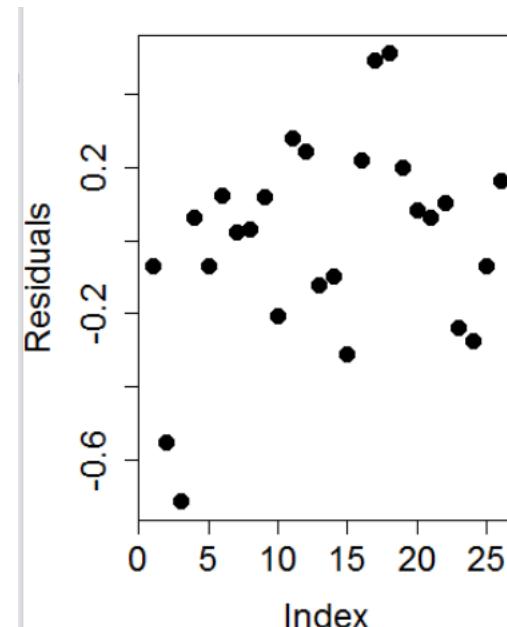


Standardized Residuals

- Checks for influence/outliers in the y-direction
- Rule of thumb: any standardized residual greater than or equal to +/- 2 are outliers

$$e_s = \frac{e_i}{\sqrt{MSE}}$$

```
### Calculate standardized residuals (influence in y-direction)
### if residuals > +/- 2, then they're outliers
par(mfrow=c(1,2))
plot(red_lm3$res,ylab="Residuals",cex=1.5,cex.lab=1.5,cex.axis = 1.5,pch= 16)
##### extract standardized outliers--
plot(rstandard(red_lm3),ylab="StandardizedResiduals",cex=1.5,cex.lab=1.5,
      cex.axis = 1.5,pch = 16,ylim=c(-3,3))
abline(h=-2)
abline(h=2)
### Extract studentized outliers--
plot(rstudent(red_lm3),ylab="Studentized Residuals",cex=1.5,cex.lab=1.5,cex.axis =
```



Outliers

- An outlier can be deleted....But not before careful examination as to **why** the observation is flagged
 - If no obvious reason, then no logic in deleting the observation
- Is the model appropriate to begin with?
 - If a flawed model, may be inadvertently pointing to data as outliers, but are really not
- A weighted linear regression model could be tried
 - Influential observations are given less ‘weight’ in the model compared to other observations

Lunch

Exposure/Response Modeling II-A

Single Binary Variables vs Binary Response
Using Chi-Squared Tests, Relative Risk, and Odds Ratios

Types of Clinical Data Variables

- **Qualitative or Categorical** – a variable with categories
 - Ordinal – meaningful ranking or scale, but not quantified
 - Example: pain status (mild, moderate, severe)
 - Example: cancer staging or grades (1, 2, 3, or 4)
 - Nominal – NO meaningful rankings
 - Binomial – yes/no; gender (male/female), genotype (for some genes)
 - Other – blood types, race, genotype (for some genes)
- **Quantitative** – a variable with numeric values
 - Continuous – measured on a continuous scale
 - Example: age, body weight, drug concentration
 - Discrete – measured on a discrete scale
 - Example: number of seizures within a time period (count data), survival (event data)

Binomial Distributions

- Must satisfy the following conditions:
 1. Outcomes (responses) are binary: only two possible outcomes in each trial (success or failure)
 - e.g. Disease status (yes/no), Nausea (yes/no)
 2. Trials are independent: the outcome of one trial is independent of the outcome of another trial
 3. Total numbers of trials is fixed in advance (n)
 4. Success probability is the same across all the trials (p)

Binomial Distributions

- If a random variable Y is the number of successes, then Y is called the binomial random variable (parameters of n, p)

$$Y \sim Bin(n, p), \quad Y = 0, 1, \dots, n$$

$$E(Y) = n * p \qquad \text{Expected value of number of successes}$$

$$SD(Y) = \sqrt{np(1 - p)} \qquad \text{Standard deviation of number of successes}$$

- If $n=1$, the Y is called a Bernoulli random variable

Statistical Inferences for Binary Data

- Main parameter of interest: population proportion
- Hypothesis testing for two proportions, or groups
 - Chi-squared test
 - Confidence interval approach
- Association between two categorical variables (e.g. yes/no)
 - Odds ratio
 - Relative risk

Case Study

- Best way to teach and demonstrate exposure/response modeling, where the response is binomial (yes/no), is through an example
- Hypothetical phase II trial with a novel compound or placebo
- Trial collected demographic data
 - Weight (kg), race, gender
- Trial collected PK data
 - Have AUC exposure values for drug dosing at steady-state
- Trial collected endpoint safety response data
 - Occurrence of nausea (yes/no)
 - For patients given both drug and placebo

Dataset Exploration

- 150 patients
 - 50 given placebo
 - 100 given drug
 - Note: those given placebo has drug AUCs of zeros

```
> aedata<-read.csv("Nausea.csv", header=TRUE)
> view(aedata)
> head(aedata)
```

	ID	AUC	WT	isF	RACE	Gender	Nausea	TRT
1	1	0	84	1	Asian	Female	YES	Placebo
2	2	0	68	0	Other	Male	NO	Placebo
3	3	0	105	1	Caucasian	Female	YES	Placebo
4	4	0	26	0	Black	Male	NO	Placebo
5	5	0	48	1	Asian	Female	YES	Placebo
6	6	0	66	1	Asian	Female	NO	Placebo

```
> summary(aedata)
```

ID	AUC	WT	isF	RACE
Min. : 1.00	Min. : 0	Min. : 26.00	Min. : 0.00	Asian : 24
1st Qu.: 38.25	1st Qu.: 0	1st Qu.: 56.00	1st Qu.: 0.00	Black : 37
Median : 75.50	Median : 2709	Median : 67.00	Median : 1.00	Caucasian: 59
Mean : 75.50	Mean : 2968	Mean : 72.04	Mean : 0.52	Hispanic : 10
3rd Qu.: 112.75	3rd Qu.: 4766	3rd Qu.: 85.00	3rd Qu.: 1.00	Other : 20
Max. : 150.00	Max. : 13933	Max. : 145.00	Max. : 1.00	

Gender	Nausea	TRT
Female: 78	NO : 109	Drug : 100
Male : 72	YES: 41	Placebo: 50

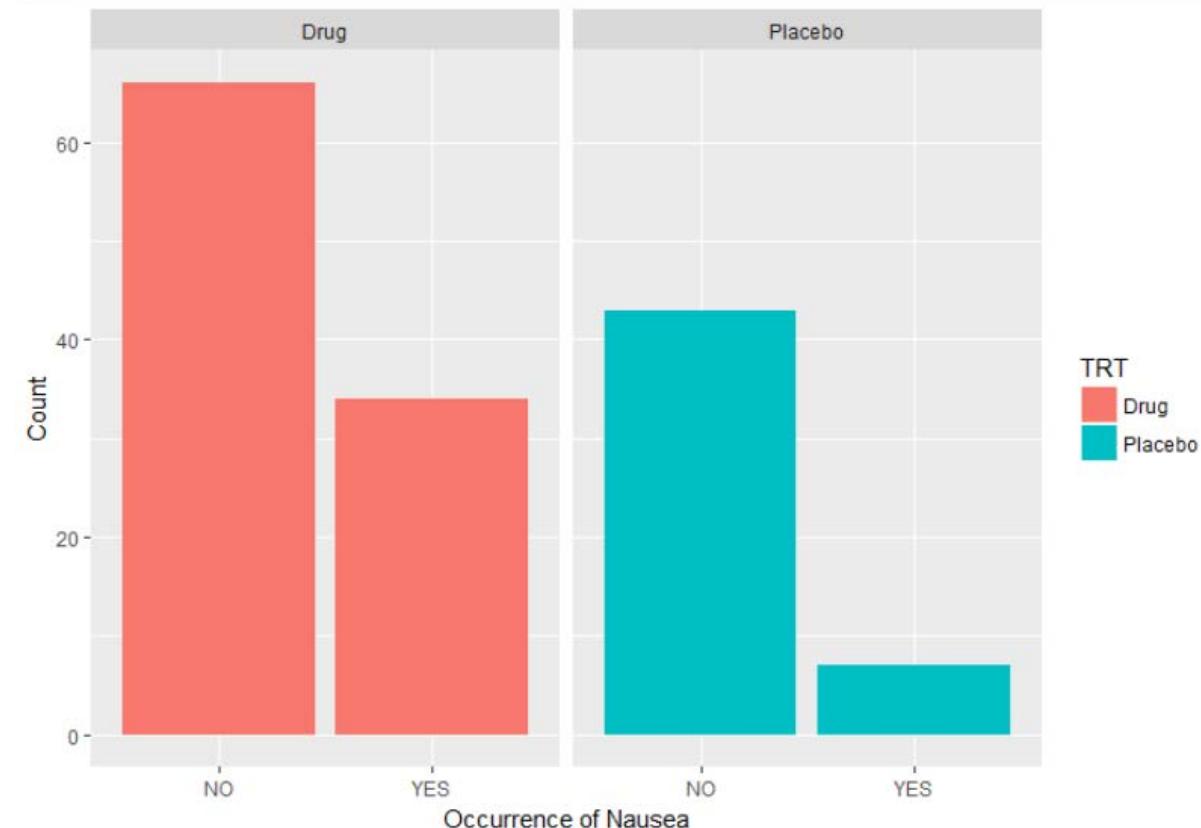
Research Questions

1. Is the risk (probability) of experiencing nausea after taking drug higher than after taking placebo?

2. Do females have a higher risk of nausea over males?

Exploratory Plots

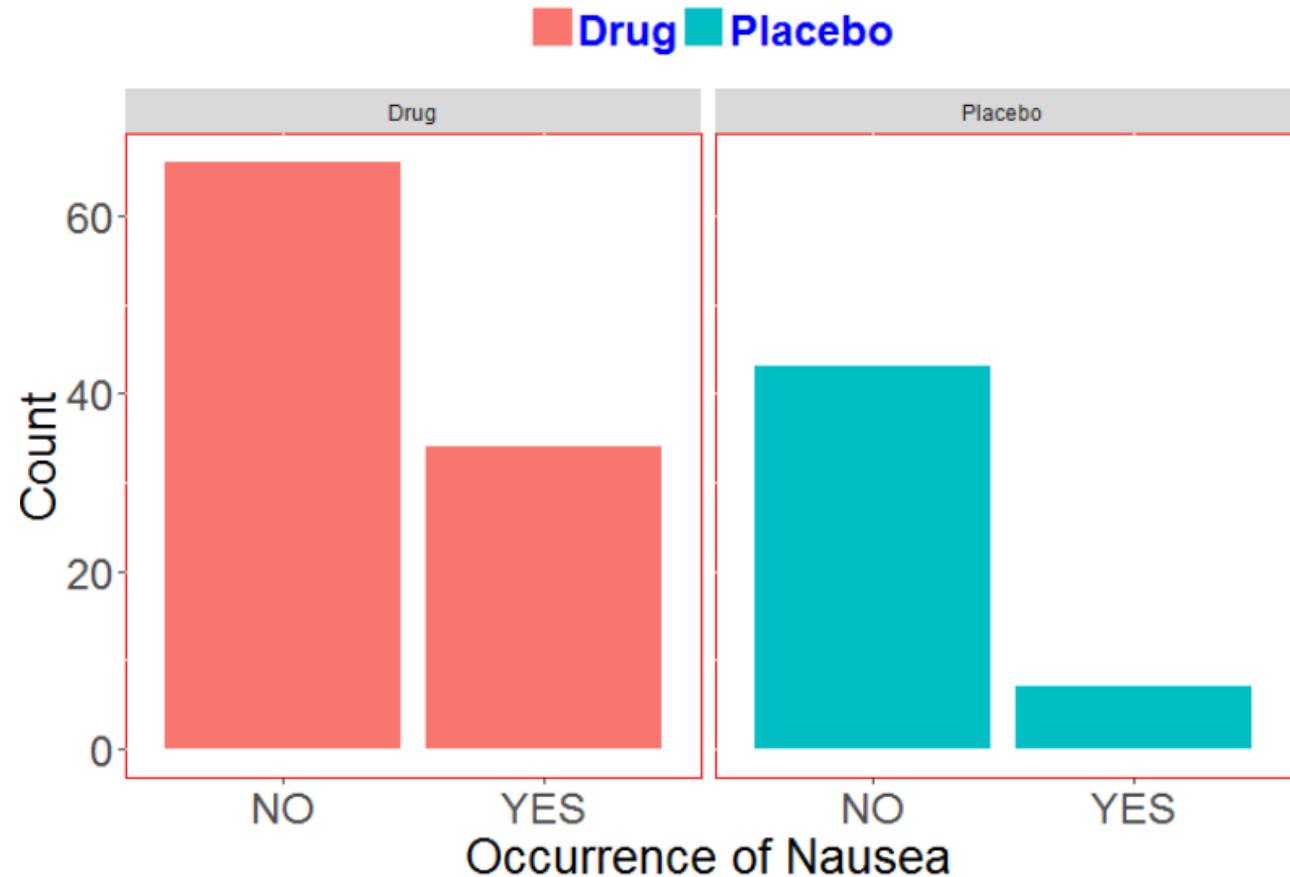
- Use “ggplot2” to make nicer looking, more complex plots
- “Counts” the number of instances of a response
 - Can stratify based on treatment



```
### Bar graphs ###
library(ggplot2)
plot1<-ggplot(data=aedata, aes(x=factor(Nausea), fill=TRT)) +
  geom_bar(stat="count") +
  facet_grid(.~ TRT) +
  xlab("Occurrence of Nausea") +
  ylab("Count")
plot1
```

Exploratory Plots

- Can make the ggplot look even nicer with some plot attributes that can be tailored to your liking



```
#### plot attributes - axis labels, axis titles, background, legend options ####
plot1+ theme(panel.background = element_rect(fill='white', colour='red'))+
  theme(axis.title.y = element_text(colour = 'black', size = 20))+
  theme(axis.text.y = element_text(size = 18))+
  theme(axis.title.x = element_text(colour = 'black', size = 20))+
  theme(axis.text.x = element_text(size = 18))+
  theme(plot.title = element_text(lineheight=.8, face="bold"))+
  theme(legend.position="top")+
  theme(legend.title=element_blank())+
  theme(legend.text = element_text(colour="blue", size = 18, face = "bold"))
```

Early Observations

- To get an idea of where the data may lead, can calculate probability of drug causing nausea and compare that to placebo
- 2 x 2 contingency table
- Utilizes probability...

$$P(\text{Nausea} = \text{No} | \text{Drug}) = \frac{P(\text{Nausea} = \text{No} \& \text{Drug})}{P(\text{Drug})}$$

```
> table(Nausea=aedata$Nausea,Treatment=aedata$TRT)
```

Nausea	Treatment		109
	Drug	Placebo	
NO	66	43	109
YES	34	7	41
	100	50	150

$$P(\text{Nausea} = \text{No} | \text{Drug}) = \frac{66/150}{100/150} = \frac{66}{100} = 0.66$$

Chi-Squared Test of Proportions

- Let's put these observations to a statistical test
- Since we're dealing in binomial response data, we have to measuring response in incidences per opportunity
 - **Proportion** of incidence that occurred
- Is the risk (probability) of experiencing nausea after taking drug **different** than after taking placebo?

$$\begin{aligned} H_0: p_{\text{drug}} - p_{\text{placebo}} &= 0 \\ H_A: p_{\text{drug}} - p_{\text{placebo}} &\neq 0 \end{aligned}$$

Can only be used to explore binomial vs binomial data:

- Nausea wrt gender (male/female)
- Nausea wrt treatment (drug/placebo)

$$H_0: P(\text{Nausea} = \text{Yes} | \text{drug}) = P(\text{Nausea} = \text{Yes} | \text{Placebo})$$

Chi-Squared Test of Proportions

- Uses the Chi-square test statistic
 - O: observed
 - E: expected

$$\chi_s^2 = \sum \frac{(O - E)^2}{E} \sim \chi_{df}^2 = 1$$

Nausea	Drug	Placebo	
NO	66	43	109
YES	34	7	41
	100	50	150

Chi-Squared Test of Proportions

- Uses the Chi-square test statistic
 - O: observed
 - E: expected

$$\chi_s^2 = \sum \frac{(O - E)^2}{E} \sim \chi_{df}^2 = 1$$

	Drug	Placebo	Totals
Nausea = No	66 E1 = (72.66)	43 E2 = (36.33)	109
Nausea = Yes	34 E3 = (27.33)	7 E4 = (13.66)	41
Totals	100	50	150

$$E_n = \frac{\text{row total} * \text{column total}}{\text{grand total}}$$

If “observed” is much different than “expected”, will result in a high value for chi-square statistic

*This can be repeated for any other variable

Chi-Squared Test of Proportions

- Run a chi-squared test
- Chi-squared statistic is 5.7437
 - With one degree of freedom
 - Df = (#rows – 1) * (# columns – 1)
- Based on df=1, the chi-squared distribution, and assuming a type-I error rate (α) of 0.05, the critical value is 3.84
 - Chi-sq statistic value > 3.84 is significant ($p<0.05$)
 - Chi-sq statistic value < 3.84 is NOT significant ($p>0.05$)

```
### Chi-squared test--Nausea by TRT--  
ae.test<-chisq.test(aedata$Nausea, aedata$TRT)  
### Display test results---  
ae.test  
  
Pearson's Chi-squared test with Yates' continuity correction  
  
data: aedata$Nausea and aedata$TRT  
X-squared = 5.7437, df = 1, p-value = 0.01655
```

Can determine the critical value based on degrees of freedom and alpha

```
> qchisq(0.95,1)  
[1] 3.841459
```

5.7437 > 3.84, thus reject null.

P=0.01655

Observed data has sufficient evidence to conclude that occurrence of nausea is **different** between drug and placebo groups

Confidence Interval Approach

- Can compare proportions of occurrences in each group
- Use confidence intervals to determine if significantly different
 - Instead of chi-squared test
 - Same result, different path to get there
- Typically use 95% CIs
 - $100(1-\alpha)$, where $\alpha=0.05$
- Obtain CIs for the *difference* in the two proportions
- If CI contains zero, then accept null hypothesis (H_0 , no difference in occurrence of nausea based on group)
- If CI does NOT contain zero, then reject H_0

$$(\hat{p}_1 - \hat{p}_2) \pm [Z_{1-\alpha/2} SE_{(\hat{p}_1 - \hat{p}_2)} + \frac{1}{2}(\frac{1}{n_1} + \frac{1}{n_2})]$$
$$\hat{p}_1 = \frac{y_1}{n_1}; \hat{p}_2 = \frac{y_2}{n_2}$$
$$SE_{\hat{p}_1 - \hat{p}_2} = \sqrt{\frac{\hat{p}_1(1 - \hat{p}_1)}{n_1} + \frac{\hat{p}_2(1 - \hat{p}_2)}{n_2}}$$

Confidence Interval Approach

- Test is based on observations deemed a “success”
 - i.e. those experiencing nausea (“Yes”)
 - A “failure” would be no nausea, or a failure to observe the response

Nausea	Drug	Placebo	
NO	66	43	109
YES	34	7	41
	100	50	150

- 95% CI does NOT contain zero: reject Null

```
> prop.test(c(34,7), c(100,50))
```

2-sample test for equality of proportions with continuity correction

```
data: c(34, 7) out of c(100, 50)
X-squared = 5.7437, df = 1, p-value = 0.01655
alternative hypothesis: two.sided
95 percent confidence interval:
 0.05131954 0.34868046
sample estimates:
prop 1 prop 2
 0.34    0.14
```

Relative Risk

- Another way to test if there's a difference in probabilities of some event (response) occurring between two groups
- Measure of dependence for two nominal variables
- Ratio of probabilities:

$$RR = \frac{P_1}{P_2} = \frac{P(Nauseau = Yes | Female)}{P(Nausea = Yes | Male)}$$

- RR = 1 means that there's equal probability of an event occurring in either group
 - Implies independence between the two nominal variables

Relative Risk

- Can only compare **two** groups at a time
- Let's try 1) occurrence of nausea and 2) gender
 - We'll have to pick either subjects on drug or placebo, so we'll obviously pick drug group
 - Need to subset those given drug from dataset

```
### Subset Treatment by Drug--  
aedata_drug<-aedata[aedata$TRT=="Drug",]
```

> `table(aedata_drug$Nausea, aedata_drug$Gender)`

	Female	Male	
NO	24	42	109
YES	28	6	41
	100	50	150

- Sample relative risk (SRR), an inference of population RR:

$$RR = \frac{P_1}{P_2} = \frac{P(Nauseau = Yes | Female)}{P(Nausea = Yes | Male)} = \frac{28/52}{6/48} = 4.3$$

*Risk of nausea is 4.3x greater in females vs males

Relative Risk

- Calculate a 95% CI for the SRR estimate

```
> gentable<-table(aedata_drug$Nausea, aedata_drug$Gender)
> gentable<-table(aedata_drug$Nausea, aedata_drug$Gender)
> col.total_F<-gentable[1,1]+gentable[2,1]
> col.total_M<-gentable[1,2]+gentable[2,2]
> p_female<-gentable[2,1]/col.total_F
> p_male<-gentable[2,2]/col.total_M
> RR<-p_female/p_male
> RR
[1] 4.307692 1.955697 9.488288
```

```
#### C.I for log(RR)---
SE_logRR<-sqrt(((1-p_female)/(p_female*col.total_F)+  

                  (1-p_male)/(p_male*col.total_M)))
logRR<-log(p_female/p_male)
ci.lb<-logRR-1.96*SE_logRR
ci.ub<-logRR+1.96*SE_logRR
#### Exponentiate to get CI on normal scale-----
ci_RR<-c(RR,exp(ci.lb),exp(ci.ub))
ci_RR
```

Odds Ratio

- Another way to test if there's a difference in probabilities of some event (response) occurring between two groups
- An odds ratio (OR) is the ratio of the odds of an event in either of two groups

$$OR = \frac{p_1 / (1 - p_1)}{p_2 / (1 - p_2)}$$
$$\frac{P_1}{P_2} = \frac{P(Nauseau = Yes | Female)}{P(Nausea = Yes | Male)}$$

```
OR = (p_female/(1-p_female))/(p_male/(1-p_male))
```

```
OR
```

```
> OR
```

```
[1] 8.166667
```

```
> table(aedata_drug$Nausea, aedata_drug$Gender)
```

	Female	Male	
NO	24	42	66
YES	28	6	34
	52	48	100

Females are 8.17x more likely to experience nausea
vs males

Confidence Interval for Odds Ratio (OR)

- Calculating CI of the OR is based on normal approximation with respect to log OR

- can't have a negative OR
- OR is **not** normally distributed
- OR is **log**-normally distributed

$$\log(OR) \pm Z_{\alpha/2} * SE_{\log(OR)}$$

$$\text{where } SE_{\log(OR)} = \sqrt{\left\{ \frac{1}{n_{11}} + \frac{1}{n_{12}} + \frac{1}{n_{21}} + \frac{1}{n_{22}} \right\}}$$

- Then transform log CI to normal scale

- 100(1- α)% CI for log OR

```
log_OR
##### C.I for OR ---
SE_logOR<-sqrt((1/gentable[1,1])+(1/gentable[1,2])+(1/gentable[2,1]) +
(1/gentable[2,2]))
ci.lb<-log_OR - 1.96 * SE_logOR
ci.ub<-log_OR + 1.96 * SE_logOR
#### Exponentiate to get CI on normal scale---
ci_OR<-c(OR,exp(ci.lb),exp(ci.ub))
ci_OR
```

n11=row1, column1 n21=row2, column1
n12=row1, column2 n22=row2, column2

	Female	Male
NO	24	42
YES	28	6

```
> ci_OR
[1] 8.166667 2.961407 22.521203
>
```

Relative Risk and Odds Ratio

- Interchangeable with a conversion factor

$$OR = RR * \left(\frac{1 - p_2}{1 - p_1} \right)$$

PM Break

Exposure/Response Modeling II-B

Continuous or Discrete Variables vs Binary Response
Using Logistic Regression

Case Study

- Hypothetical phase II trial with a novel compound or placebo
- Trial collected demographic data
 - Weight (kg), race, gender
- Trial collected PK data
 - Have AUC exposure values for drug dosing at steady-state
- Trial collected endpoint safety response data
 - Occurrence of nausea (yes/no)
 - For patients given both drug and placebo

Research Questions

1. Is the risk (probability) of experiencing nausea after taking drug higher than after taking placebo?
 - Yes, answered with a chi-squared test
2. Do females have a higher risk of nausea over males?
 - Yes, answered with relative risk and odds ratio
3. Are occurrences of nausea related to drug exposure (AUC)?
 - Since exposure (AUC) is a continuous variable, can we use a linear regression model?

Assumptions for OLS

- ϵ_i are normal random variables
 - $N(0, \sigma^2)$ is a constant variance, i.e. homogenous variance
 - Violated, because response variable is binary, not continuous with normal distribution. Thus residual error for response is not normally distributed with constant variance
- Because response is binary (yes/no), the residual error variance is a function of its probability
 - $\text{Var}(\epsilon_i) = P_i (1-P_i)$
 - As probabilities change with predictor variable (prob of response different for males vs females within gender variable), so do the errors associated with each predictor variable
 - Violate assumption of homogenous variance

Logistic Regression

- A way to model the **mean response** against predictor variable(s) through a logistic (or a logit) function
 - No “residual” error like in linear regression for continuous variables (obs-pred)
 - In logistic regression, error: $\text{Var}(\varepsilon_i) = P_i * (1-P_i)$
 - Error is a function of probability of having success or no success
- *Generalized Linear Model (GLM)*
 - 3 components:
 1. **Distributional assumption** (random component)
 - Identifies response (Y) and its probability distribution
 2. **Systematic component**
 - Specifies predictor variables used in a linear predictor function
$$g[E(Y)] = \beta_0 + \beta_1 x_1 + \beta_2 x_2$$
 3. **Link function**
 - Specifies $g(\cdot)$, links the random (1) and systematic (2) components together

Generalized Linear Model

1) Distributional Assumption for Response (Y)

- Exponential family

$$f(Y : \mu) = \exp[a(Y)*b(\mu) + c(\mu) + d(Y)]$$

μ : parameter (not population mean)

b: natural parameter of the distribution

- Binomial distribution: a member of the exponential family

$$f(Y : p) = \binom{n}{Y} p^Y (1 - p)^{n-Y}$$

*If $n=1$, Y follows a Bernoulli distribution

- for pos/neg responses

$$b(\mu) = \log\left(\frac{\mu}{1-\mu}\right) = \log\left(\frac{p}{1-p}\right)$$

b: natural parameter of *binomial* distribution

Generalized Linear Model

2) Systematic Component

- Relates effect of predictor variables (covariates) to the *transformed mean response* through a linear model

$$g[E(Y)] = \beta_0 + \beta_1 x_1 + \beta_2 x_2$$

3) Link Function

- Links the random component (distributional assumption) to the systematic component
- the logit function for a Bernoulli distribution

$\text{logit}(P(Y=1))$: logit function for probability of success (event occurred)

Logit Function

- $\text{logit}(P(Y=1))$
- $g[E(Y)] = \beta_0 + \beta_1 x_1 + \beta_2 x_2$

**General form of a logistic regression with
2 predictor variables**

- $\log\left(\frac{p(Y=1)}{1-p(Y=1)}\right) = \beta_0 + \beta_1 x_1 + \beta_2 x_2$



systematic component

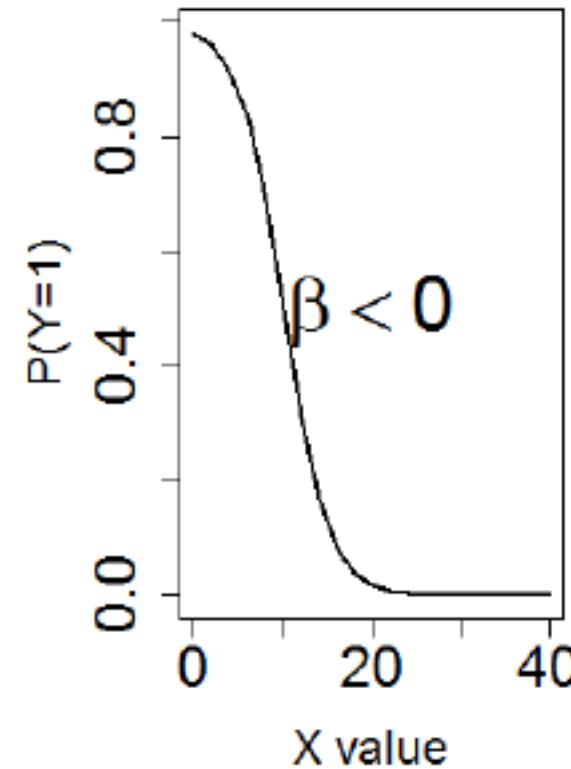
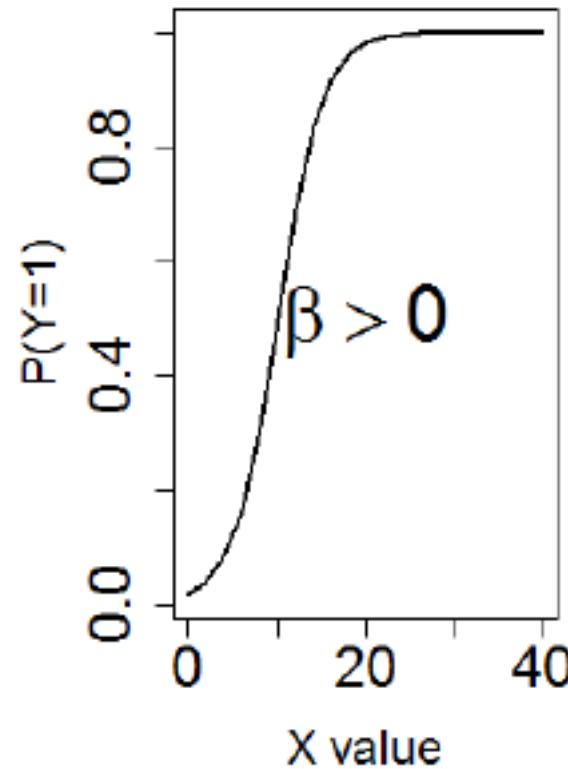
random component (distributional assumption)

Exponentiate both sides of logit function, solve for $P(Y=1)$ to get the probability of the response occurring:

$$P(Y = 1) = \frac{\exp(\beta_0 + \beta_1 x_1 + \beta_2 x_2)}{1 + \exp(\beta_0 + \beta_1 x_1 + \beta_2 x_2)}$$

Shape of Logistic Regression Functions

- Curvilinear relationship of the response function
- When slope (β) of predictor variable (x) is positive, probability curve goes from 0 to 1
- When slope (β) of predictor variable (x) is negative, probability curve goes from 1 to 0



Logistic Regression of Nausea Data in R

- Run the intercept-only model first (no predictor variables; “base” model)
 - i.e. models the placebo effect

$$\log\left(\frac{p(Y=1)}{1-p(Y=1)}\right) = \beta_0$$

- Deviance: a measure of a model’s goodness of fit

- Calculated by the difference in the log-likelihood of a “saturated” model (with all available covariates included) and the current model

$$G^2 = 2[LL(\beta_{MAX}|Y) - LL(\beta|Y)]$$

$$G^2 \sim \chi^2_{df=N-p}$$

Deviance (G^2) follows a chi-squared distribution, with degrees of freedom equal to the # of observations - # parameters.
A “saturated” model is one where $N=p$.

```
> fit.nausea<-glm(aedata_drug$Nausea ~ 1, family=binomial)
> summary(fit.nausea)

Call:
glm(formula = aedata_drug$Nausea ~ 1, family = binomial)

Deviance Residuals:
    Min      1Q  Median      3Q     Max 
-0.9116 -0.9116 -0.9116  1.4689  1.4689 

Coefficients:
            Estimate Std. Error z value Pr(>|z|)    
(Intercept) -0.6633    0.2111 -3.142   0.00168 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

Null deviance: 128.21 on 99 degrees of freedom
Residual deviance: 128.21 on 99 degrees of freedom
AIC: 130.21

Number of Fisher Scoring iterations: 4
```

$$G^2 = 2[LL(\beta_{MAX}|Y) - LL(\beta|Y)]$$

Deviance

$$G^2 \sim \chi^2_{df=N-p}$$

- The saturated model is considered to be the “ideal” model
- Perform chi-squared test to compare proportions of the “ideal” model with the current model
- In this case, we want $p>0.05$ (accept null hypothesis)
 - Meaning the current model is at least as good as the “ideal” model
 - Meaning the current model has adequate fit to the data (no worse than “ideal” model)
 - If $p<0.05$, then current model is significantly different from ideal model, and only in a “worse-fit” manner
 - Can’t be “better” than the “ideal” model

Full Model

- Include all available predictor variables/covariates

$$\log\left(\frac{p(Y=1)}{1-p(Y=1)}\right) = \beta_0 + \beta_1 \text{AUC} + \beta_2 \text{WT} + \beta_3 \text{Gender} + \beta_4 \text{Race}$$

- Full model identified Gender and AUC as potential significant predictors
- Race and body weight were not significant predictors
- Null deviance remains 128.21 (“ideal” model)
 - df=99 (100 obs – 1 parameter(intercept))
- Current model’s deviance (residual) is 97.69
 - df=92 (100 obs – 8 parameters)

```
Call:  
glm(formula = Nausea ~ AUC + WT + relevel(Gender, "Male") + RACE,  
     family = binomial, data = aedata_drug)  
  
Deviance Residuals:  
    Min      1Q   Median      3Q      Max  
-1.7298 -0.7318 -0.3999  0.8690  2.2708  
  
Coefficients:  
              Estimate Std. Error z value Pr(>|z|)  
(Intercept) -2.5176428  1.2499799 -2.014  0.0440 *  
AUC          0.0002526  0.0001209  2.088  0.0368 *  
WT          -0.0028155  0.0116878 -0.241  0.8096  
relevel(Gender, "Male")Female 2.2987503  0.5653896  4.066 4.79e-05 ***  
RACEBlack   -0.0569957  0.7927891 -0.072  0.9427  
RACECaucasian -0.7233218  0.7563302 -0.956  0.3389  
RACEHispanic -1.7533496  1.3011010 -1.348  0.1778  
RACEOther   -0.8833187  1.0475194 -0.843  0.3991  
---  
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1  
  
(Dispersion parameter for binomial family taken to be 1)  
  
Null deviance: 128.21 on 99 degrees of freedom  
Residual deviance: 97.69 on 92 degrees of freedom  
AIC: 113.69  
  
Number of Fisher Scoring iterations: 5
```

Saturated (Full) Model

- Is the full model a statistically-significant better fit to the data than the base model?
- Test model improvement using deviance using chi-squared test

```
(Dispersion parameter for binomial family taken to be 1)
```

```
Null deviance: 128.21 on 99 degrees of freedom  
Residual deviance: 97.69 on 92 degrees of freedom  
AIC: 113.69
```

```
Number of Fisher Scoring iterations: 5
```

```
> pchisq(deviance(fit.nausea.full),df.residual(fit.nausea.full),lower=FALSE)  
[1] 0.3227307  
>
```

- $p=0.322$, thus $p>0.05$, so accept H_0 that current full model is as good as the “ideal” model

Other Model Evaluations

- Akaike information criterion (AIC)

- a measure of a model's goodness of fit
- the lower value means the better the model fit
- Full model has a better fit than base
 - Inclusion of predictor variables significantly improves model's ability to predict the data

$$AIC = -2 * \text{log-likelihood} + 2p$$

Base Model

```
Call:
glm(formula = aedata_drug$Nausea ~ 1, family = binomial)

Deviance Residuals:
    Min      1Q  Median      3Q     Max 
-0.9116 -0.9116 -0.9116  1.4689  1.4689 

Coefficients:
            Estimate Std. Error z value Pr(>|z|)    
(Intercept) -0.6633    0.2111 -3.142   0.00168 **  
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 128.21 on 99 degrees of freedom
Residual deviance: 128.21 on 99 degrees of freedom
AIC: 130.21
```

Number of Fisher Scoring iterations: 4

Full Model

```
Coefficients:
            Estimate Std. Error z value Pr(>|z|)    
(Intercept) -2.5176428  1.2499799 -2.014  0.0440 *  
AUC          0.0002526  0.0001209  2.088  0.0368 *  
WT           -0.0028155  0.0116878 -0.241  0.8096    
relevel(Gender, "Male")Female  2.2987503  0.5653896  4.066 4.79e-05 *** 
RACEBlack    -0.0569957  0.7927891 -0.072  0.9427    
RACECaucasian -0.7233218  0.7563302 -0.956  0.3389    
RACEHispanic  -1.7533496  1.3011010 -1.348  0.1778    
RACEOther     -0.8833187  1.0475194 -0.843  0.3991    
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 128.21 on 99 degrees of freedom
Residual deviance:  97.69 on 92 degrees of freedom
AIC: 113.69
```

Number of Fisher Scoring iterations: 5

Reduced Model

- Since AUC and Gender were significant, and WT and Race non-significant, predictors in the full model; remove WT and Race from model
- “reduced” model
- all predictor variables are significant
- AIC = 107.33 (lower than full model’s 113.69)
 - better fit than full
- Deviance = 101.33
 - > pchisq(deviance(fit.nausea.reduced),df.residual(fit.nausea.reduced),lower=FALSE)
[1] 0.3616901
- p=0.36, accept H_0
- reduced model as good as “ideal”

```
> fit.nausea.reduced<-glm(formula = Nausea ~ AUC + relevel(Gender, "Male"), family = binomial,
+   data = aedata_drug)
> summary(fit.nausea.reduced)

Call:
glm(formula = Nausea ~ AUC + relevel(Gender, "Male"), family = binomial,
     data = aedata_drug)

Deviance Residuals:
    Min      1Q  Median      3Q     Max 
-1.5296 -0.8732 -0.4013  0.9067  2.2141 

Coefficients:
              Estimate Std. Error z value Pr(>|z|)    
(Intercept) -3.2905854  0.7642531 -4.306 1.67e-05 ***
AUC          0.0002803  0.0001174  2.388  0.0169 *  
relevel(Gender, "Male")Female 2.1970295  0.5455977  4.027 5.65e-05 *** 
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

Null deviance: 128.21 on 99 degrees of freedom
Residual deviance: 101.33 on 97 degrees of freedom
AIC: 107.33

Number of Fisher Scoring iterations: 4
```

Interpretation of the Model

- Reduced model is the Final model

$$\log\left(\frac{p(Y=1)}{1-p(Y=1)}\right) = -3.29 + 0.00028(\text{AUC}) + 2.19(\text{Gender})$$

- For a fixed AUC, *exponentiating* the coefficient of gender would give the odds ratio for occurrence of nausea in females

```
> exp(2.19)
[1] 8.935213
```

- OR = 8.94, meaning females 8.94x more likely to occur as of that in males *at a given AUC value*
 - comparable to the 8.19 OR for females vs males calculated from univariate analysis

Interpretation of the Model

- Reduced model is the Final model

$$\log\left(\frac{p(Y=1)}{1-p(Y=1)}\right) = -3.29 + 0.00028(\text{AUC}) + 2.19(\text{Gender})$$

- For the continuous covariate AUC: for a 1000-unit change in AUC from zero, the odds ratio is

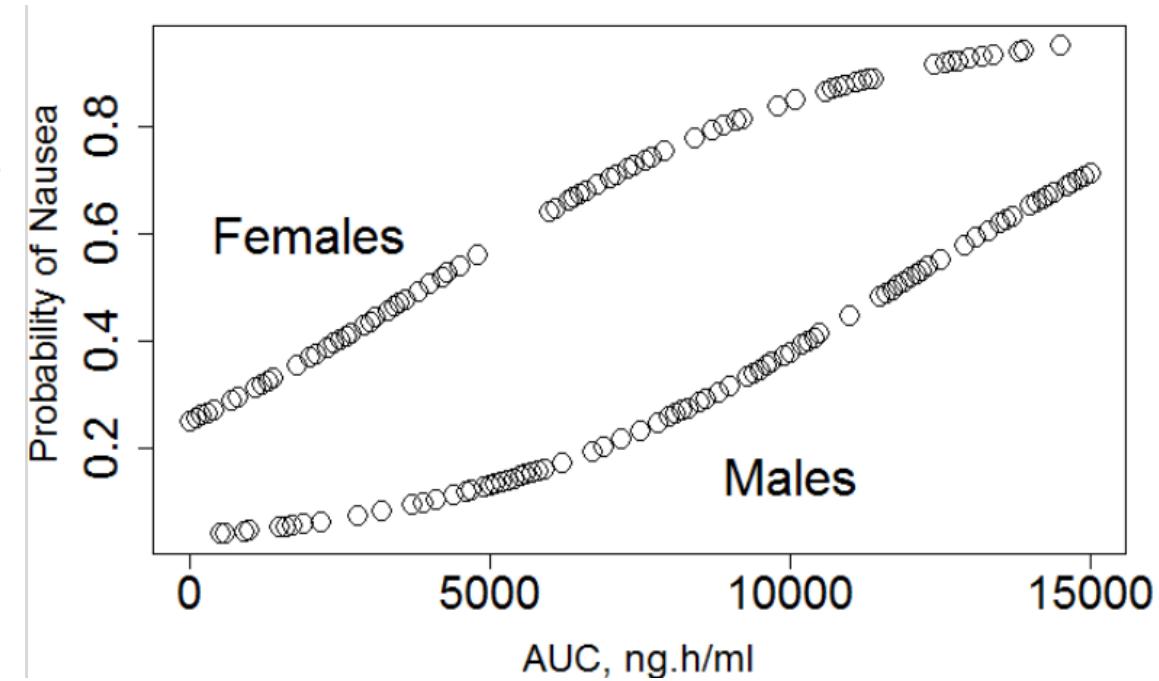
```
> exp(0.00028*1000)
[1] 1.32313
```

Depiction of Logistic Regression

$$\log\left(\frac{p(Y=1)}{1-p(Y=1)}\right) = -3.29 + 0.0028(\text{AUC}) + 2.19(\text{Gender})$$

- Let's plot the *probability of nausea* with increasing AUC based on the odds ratio, for both males and females

```
install.packages("faraway")
library(faraway)
par(mfrow=c(1,1))
x1_AUC<-seq(0,15000, by = 100)
x2_gender<-rbinom(length(x1_AUC), 1,0.5)
fit.logit<-ilogit(fit.nausea.reduced$coef[1]+fit.nausea.reduced$coef[2]*x1_AUC+
    fit.nausea.reduced$coef[3]*x2_gender)
plot(x1_AUC, fit.logit,xlab="AUC, ng.h/ml",
      ylab="Probability of Nausea",cex=1.8, cex.lab=1.5,cex.axis = 1.8)
text(2000,0.6, "Females", cex=2.0)
text(10000,0.15,"Males" , cex=2.0)
```



Confidence Intervals for the OR

- The parameter (slope) *estimate* for each variable is the log odds ratio
- Because an *estimate*, there's uncertainty about whether that value is the “true population” value
- To assess the extent of the uncertainty around that parameter estimate, calculate the 95% CI

```
### 95% CIs for odds ratio (OR)--
install.packages("MASS")
library(MASS)
exp(confint(fit.nausea.reduced))
```

```
> exp(confint(fit.nausea.reduced))
Waiting for profiling to be done...  
  
PE  
2.5 % 97.5 %  
(Intercept) 0.007026951 0.1449016 0.03725  
AUC 1.000063543 1.0005264 1.00028  
relevel(Gender, "Male")Female 3.289073483 28.7038508 8.93
```

Additional Model Diagnostics

- In addition to deviance (G^2), AIC, and confidence intervals, can use **Pearson's chi-square (χ^2) statistic**
 - Analogous to the residual sum of squares used in linear regression
 - Should be close to the deviance for adequate fit of the model to the data

```
> pearson_chisq<-sum(residuals(fit.nausea.reduced, type="pearson")^2)
> pearson_chisq
[1] 98.76349
Call:
glm(formula = Nausea ~ AUC + relevel(Gender, "Male"), family = binomial,
     data = aedata_drug)

Deviance Residuals:
    Min      1Q  Median      3Q     Max 
-1.5296 -0.8732 -0.4013  0.9067  2.2141 

Coefficients:
                                         Estimate Std. Error z value Pr(>|z|)    
(Intercept)                      -3.2905854  0.7642531 -4.306 1.67e-05 ***
AUC                           0.0002803  0.0001174  2.388  0.0169 *  
relevel(Gender, "Male")Female  2.1970295  0.5455977  4.027 5.65e-05 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

Null deviance: 128.21 on 99 degrees of freedom
Residual deviance: 101.33 on 97 degrees of freedom
AIC: 107.33

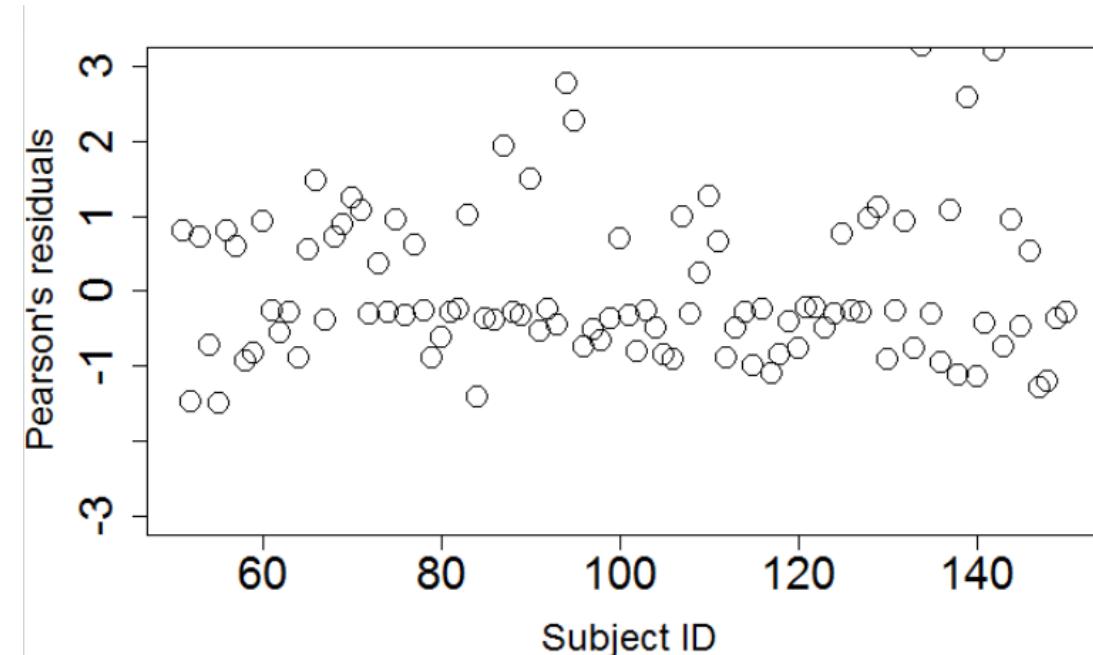
Number of Fisher Scoring iterations: 4
```

Additional Model Diagnostics

- In addition to deviance (G^2), AIC, and confidence intervals, and Pearson's chi-square (χ^2) statistic, can use **Pearson's residuals**
 - Assumed that Pearson's residuals are normally distributed
 - No obvious trends in residuals that would suspect model mis-specification

```
plot(aedata_drug$ID, residuals(fit.nausea.reduced, type="pearson"),
      xlab="Subject ID", ylab="Pearson's residuals", cex=2.0, ylim=c(-3,3),
      cex.lab=1.5, cex.axis=1.8)
```

- No trends in residuals in the 100 subjects that received drug are evident
 - Subject ID's may range from 1-150, but there are only n=100 dots



Research Questions

1. Is the risk (probability) of experiencing nausea after taking drug higher than after taking placebo?
 - Yes, answered with a chi-squared test
2. Do females have a higher risk of nausea over males?
 - Yes, answered with relative risk and odds ratio
3. Are occurrences of nausea related to drug exposure (AUC)?
 - Yes, answered with logistic regression
 - Odds ratio increases 1.3x for each 1000-fold unit increase in AUC

Day 2 Summary

- Introduction to statistical hypothesis testing in clinical trials
- Exposure/Response modeling, section I
 - Continuous or binary variable(s) vs **continuous response**
 - Linear regression: simple (one variable) and multiple (2+ variables)
- Exposure/Response modeling, section II
 - Binary variable vs **Binary response**
 - Chi-squared tests, relative risk, odds ratios
 - Binary or continuous variable(s) vs **Binary response**
 - Logistic regression

Day 3

9:00 - 9:30am:

- Day 2 recap, overflow

9:30-10:30am:

- Exposure/Response modeling III-A
 - Ordinal data
 - Proportional Odds Model

10:30-10:45am:

- Break

10:45 – 12:00pm:

- Exposure/Response modeling III-A (cont)

12:00 – 1:00pm:

- Lunch break

1:00 – 2:30pm:

- Exposure/Response modeling III-B
 - Count (longitudinal) data
 - GEE and GLMM

2:30 - 2:45pm:

- Break

2:45 – 4:30pm:

- Exposure/Response modeling III-B (cont)

Exposure/Response Modeling III-A

Continuous or Discrete Variables vs Ordinal (Discrete) Responses
Using Proportional Odds Models

Types of Clinical Data Variables

- **Qualitative or Categorical** – a variable with categories
 - **Ordinal** – meaningful ranking or scale, but not quantified
 - Example: pain status (mild, moderate, severe)
 - Example: cancer staging or grades (1, 2, 3, or 4)
 - **Nominal** – NO meaningful rankings
 - Binomial – yes/no; gender (male/female), genotype (for some genes)
 - Other – blood types, race, genotype (for some genes)
- **Quantitative** – a variable with numeric values
 - **Continuous** – measured on a continuous scale
 - Example: age, body weight, drug concentration
 - **Discrete** – measured on a discrete scale
 - Example: number of seizures within a time period (count data), survival (event data)

Examples of Ordinal Outcomes/Responses

- Adverse events
 - Dizziness (none, mild, moderate, severe)
- Therapeutic events
 - Pain relief (none, mild, moderate, complete)
 - Cancer treatment (progressive disease, stable disease, partial response, complete response)
- Disease staging
 - Cancer (stage I, II, III, IV)
- Likert scale
 - Strongly disagree, disagree, neutral, agree, strongly agree

Case Study: Pain Score

- Randomized, single dose study of placebo vs drug for pain relief after dental surgery
- 100 subjects
- Placebo vs 40 mg of drug
 - Drug named “CTM_IR”
 - IR: immediate release
- Demographics: body weight, gender
- Exposure metric: drug C_{MAX}
- Response metric: pain relief index on a scale of 1-4
 - Obtained at a single time point
 - 1=no pain relief, 2=mild relief, 3=moderate relief, 4=complete pain relief

Research Question

- How is drug exposure related to pain relief?
- Need to perform exposure/response modeling
 - Can't use linear regression (response variable NOT continuous)
 - Can't use same logistic regression (response variable NOT binary/binomial)

Pain Score Dataset

- Read in data, examine header

```
> orddata<-read.csv("Painscore.csv", sep=",")  
> head(orddata)
```

	ID	painscore	ps	Cmax		TRT	isTRT	WT	GENDER
1	1	No pain relief	1	0		Placebo	0	56	Male
2	2	Moderate pain relief	3	139	40mg	CTM_IR	1	62	Female
3	3	Complete pain relief	4	0		Placebo	0	50	Male
4	4	Complete pain relief	4	129	40mg	CTM_IR	1	66	Female
5	5	No pain relief	1	0		Placebo	0	84	Female
6	6	Complete pain relief	4	78	40mg	CTM_IR	1	65	Male

- Explore data summary

```
> lapply(orddata[, c("painscore", "TRT")], table)  
$painscore
```

Complete pain relief	Moderate pain relief	No pain relief	slight pain relief
31	21	24	24

```
$TRT
```

40mg CTM_IR	Placebo
50	50

Pain Score Dataset Summary

- Explore data summary

```
> lapply(orddata[, c("painscore", "TRT")], table)  
$painscore
```

	Complete pain relief	Moderate pain relief	No pain relief	slight pain relief	
	31	21	24	24	

```
$TRT
```

40mg CTM_IR	Placebo
50	50

```
> ftable(xtabs(~painscore+TRT, data=orddata))
```

```
                  TRT 40mg CTM_IR Placebo
```

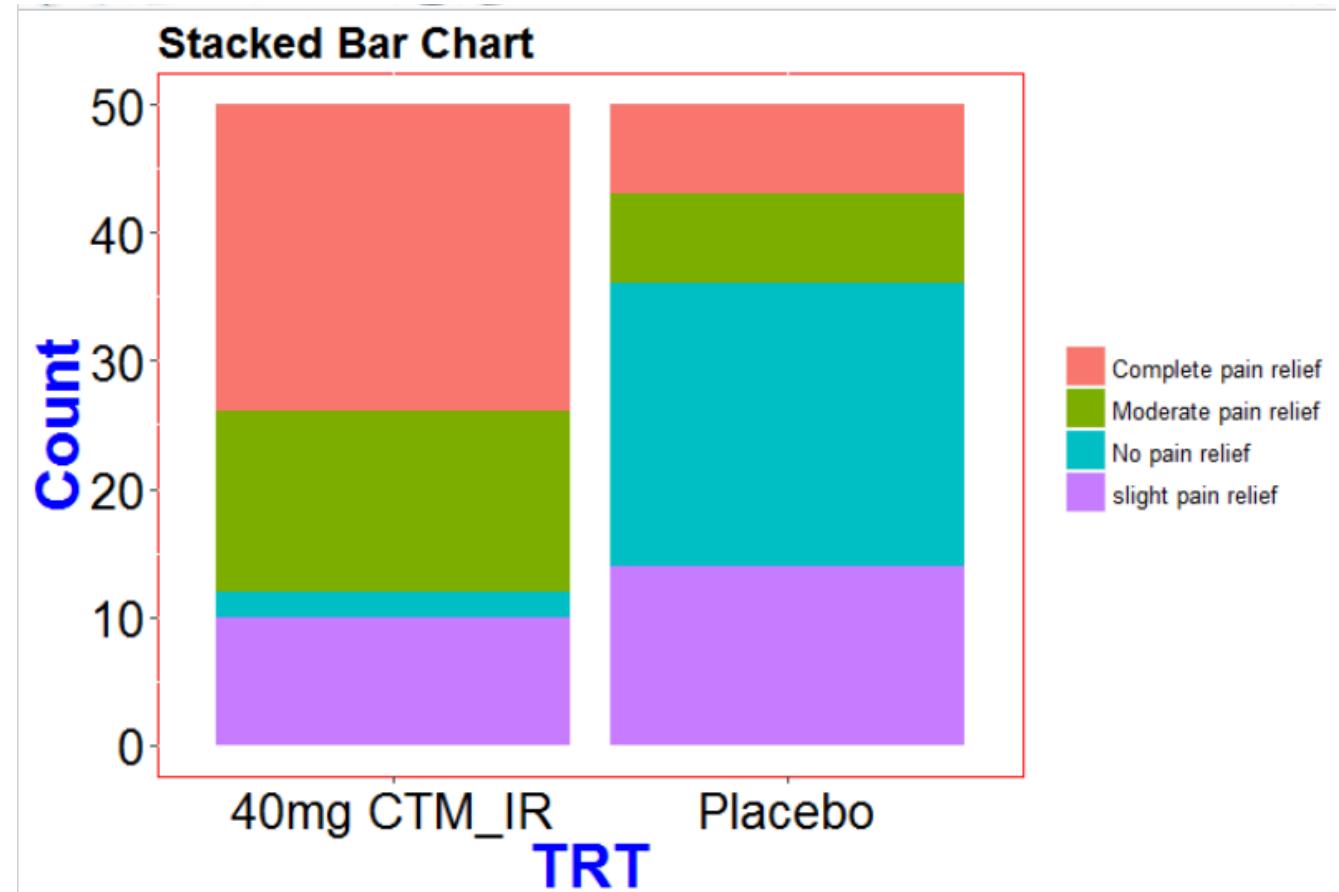
```
painscore
```

Complete pain relief	24	7
Moderate pain relief	14	7
No pain relief	2	22
slight pain relief	10	14

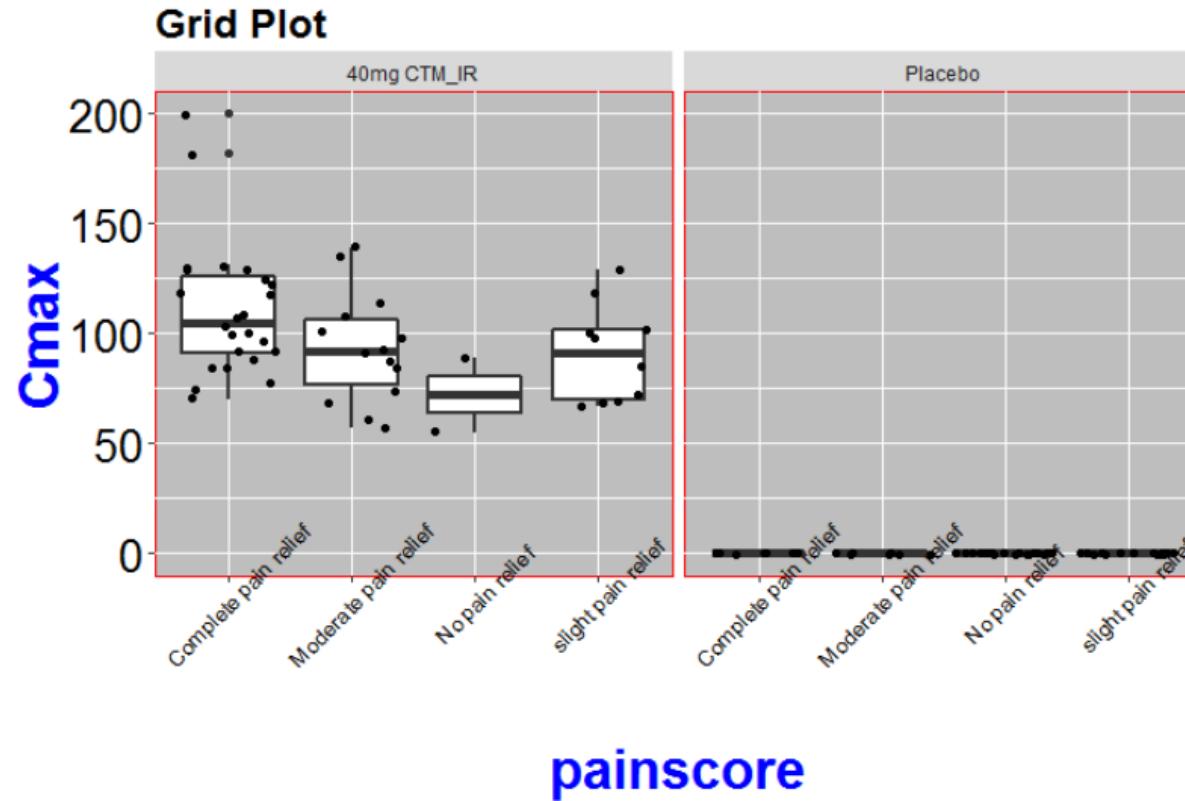
```
> |
```

Dataset Exploration

```
#### Re-arrange data for stacked bar chart--  
count<-ftable(xtabs(~painscore+TRT, data=orddata))  
install.packages("reshape2")  
library(reshape2)  
df1 = melt(count)  
library(ggplot2)  
names(df1) = c("value", "TRT", "Count")  
df1  
#### Stacked bar chart---  
plot<-ggplot(df1, aes(x=TRT, y=Count, fill=value))+  
  geom_bar(stat="identity")  
plot+ theme(panel.background = element_rect(fill='white', colour='red'))+  
  theme(axis.title.y = element_text(colour = 'blue', size = 25, face='bold'))+  
  theme(axis.text.y = element_text(size = 20, colour='black'))+  
  theme(axis.title.x = element_text(colour = 'blue', size = 25, face='bold'))+  
  theme(axis.text.x = element_text(size = 20, colour='black'))+  
  ggtitle("Stacked Bar Chart")+  
  theme(plot.title = element_text(lineheight=.8, face="bold",size = 18))+  
  theme(legend.position="right")+  
  theme(legend.title=element_blank())+  
  theme(legend.text = element_text(colour="black", size = 10))
```



Dataset Exploration – Exposure/Response



```
### side-by-side box plot (grid plot)--
plot2<-ggplot(ordrdata,aes(x=painscore,y=Cmax))+
  geom_boxplot(size=1)+
  geom_jitter(alphs=0.5)+
  facet_grid(.~TRT)
plot2+ theme(panel.background = element_rect(fill='grey', colour='red'))+
  theme(axis.title.y = element_text(colour = 'blue', size = 25, face='bold'))+
  theme(axis.text.y = element_text(size = 20, colour='black'))+
  theme(axis.title.x = element_text(colour = 'blue', size = 25, face='bold'))+
  theme(axis.text.x = element_text(size = 10, colour='black', angle=45))+
```

ggttitle("Grid Plot")+

```
theme(plot.title = element_text(lineheight=.8, face="bold",size = 18))
```

Proportional Odds Model

- Most common type of model used to analyze exposure/ordinal response data
 - Variation of logistic regression using generalized linear model
- Uses principles of cumulative logits, or logits of cumulative probabilities
 - Cumulative probability refers to the probability that the value of a random variable falls within a specified range (< or = a specified value)
 - Ex: coin flip
 - If flip coin twice, what is the probability that both coin flips would result in 0-1 heads? - need cumulative probability

Cumulative Probability

- Using coin flip example, the probability of a single coin flip is $p=0.50$ that would result in heads
- For probability of 0-1 heads from two coin flips, need to **ADD** (or cumulate) the probability that one coin flip results in 0 heads **PLUS** the probability the coin flip results in 1 heads

$$P(X \leq 1) = P(X = 0) + P(X = 1)$$

$$P(X \leq 1) = 0.25 + 0.50$$

$$P(X \leq 1) = 0.75$$

25% (1 in 4) chances that two flips result in 0 heads
50% (2 in 4) chances that two flips results in 1 heads

Cumulative Logits

- Cumulative logits account for the natural ordering in response variable
 - No pain relief, slight relief, moderate relief, complete relief
 - Uses logit function to link ordinal responses to a linear-type model
 - Similar to what binomial logistic regression models did for linking binomial responses to linear models

$$P(Y \leq j) = \pi_1 + \pi_2 + \dots + \pi_j \quad \text{Where } j = 1, 2, \dots, J$$

Cumulative Probabilities

$$\pi_1 = P(Y=1) \quad \pi_2 = P(Y=2) \quad \pi_j = P(Y=j)$$

$$\text{logit}[P(Y \leq j)] = \log \left(\frac{P(Y \leq j)}{1 - (P \leq j)} \right)$$

**Logits of
Cumulative Probabilities**

$$\text{logit}[P(Y \leq j)] = \log \left(\frac{\pi_1 + \pi_2 + \dots + \pi_j}{\pi_{j+1} + \pi_{j+2} + \dots + \pi_J} \right)$$

Proportional Odds Model

- A model that simultaneously uses all cumulative logits
- General form of the model (with n predictor variables):

$$\text{logit}[P(Y \leq j)] = \alpha_j + \beta_1 x_1 + \beta_2 + \cdots + \beta_n x_n \quad \text{Where } j = 1, 2, \dots, J$$

In our pain score example,
1 = no pain relief
2 = slight pain relief
3 = moderate pain relief
4 = complete pain relief

$$\log\left(\frac{P(Y \leq 1)}{P(Y > 1)}\right) = \alpha_1 + \beta_1 x_1$$

$$\log\left(\frac{P(Y \leq 2)}{P(Y > 2)}\right) = \alpha_2 + \beta_1 x_1$$

$$\alpha_1 < \alpha_2 < \dots < \alpha_{J-1}$$

- Each predictor variable (x) has only one slope parameter estimate (β)
 - Each response value (j) within the ordered response (Y) would have a unique intercept (α_j)

Proportional Odds Model

- General form of the model (with n predictor variables):

$$\text{logit}[P(Y \leq j)] = \alpha_j + \beta_1 x_1 + \beta_2 + \cdots + \beta_n x_n$$

- Probability that the individual is in the j^{th} category or lower is given by:

$$P(Y \leq j) = \frac{\exp(\alpha_j + \beta_1 x_1 + \beta_2 x_2 + \dots + \beta_n x_n)}{1 + \exp(\alpha_j + \beta_1 x_1 + \beta_2 x_2 + \dots + \beta_n x_n)}$$

Proportional Odds Model

- For ordinal responses, the intercept (α) will change ($\alpha_1, \alpha_2, \alpha_3, \alpha_4$), but the slope (β) of a predictor (x) will stay same
 - True for each predictor variable (x) value
- For a given ordinal response ($Y = 1, 2, 3$, or 4), can determine the change in probability when the predictor variable *value* (x) changes
 - Model called “proportional” odds model because the logits of two predictor values are *proportional* to the distance between the two values

$$\text{logit}[P(Y \leq j|x_1)] - \text{logit}[P(Y \leq j|x_2)]$$

$$= \log\left(\frac{P(Y \leq j|x_1)}{P(Y > j|x_1)}\right) = \beta_1(x_1 - x_2)$$
$$= \log\left(\frac{P(Y \leq j|x_2)}{P(Y > j|x_2)}\right)$$

Proportional Odds Model

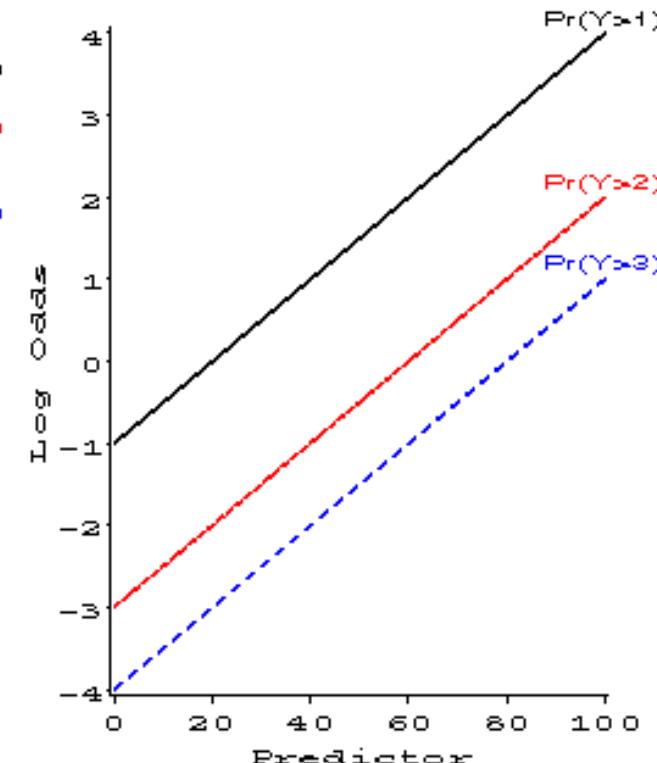
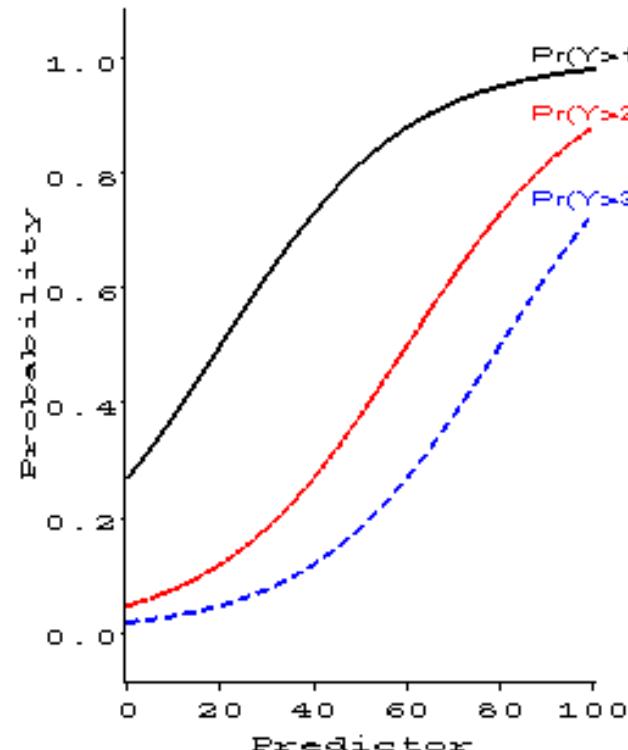
- The odds of making a response ($Y \leq j$) at $x=x_1$ are $\exp[\beta_1(x_1-x_2)]$ times the odds at $x=x_2$
- The same *proportionality* constant applies to each logit
 - E.g. $= \beta_1(x_2 - x_3)$, $= \beta_1(x_3 - x_4)$
 - E.g. $= \beta_2(x_1 - x_2)$, $= \beta_2(x_2 - x_3)$

$$\text{logit}[P(Y \leq j|x_1)] - \text{logit}[P(Y \leq j|x_2)]$$

$$= \log\left(\frac{P(Y \leq j|x_1)}{P(Y > j|x_1)}\right) = \beta_1(x_1 - x_2)$$
$$= \log\left(\frac{P(Y \leq j|x_2)}{P(Y > j|x_2)}\right)$$

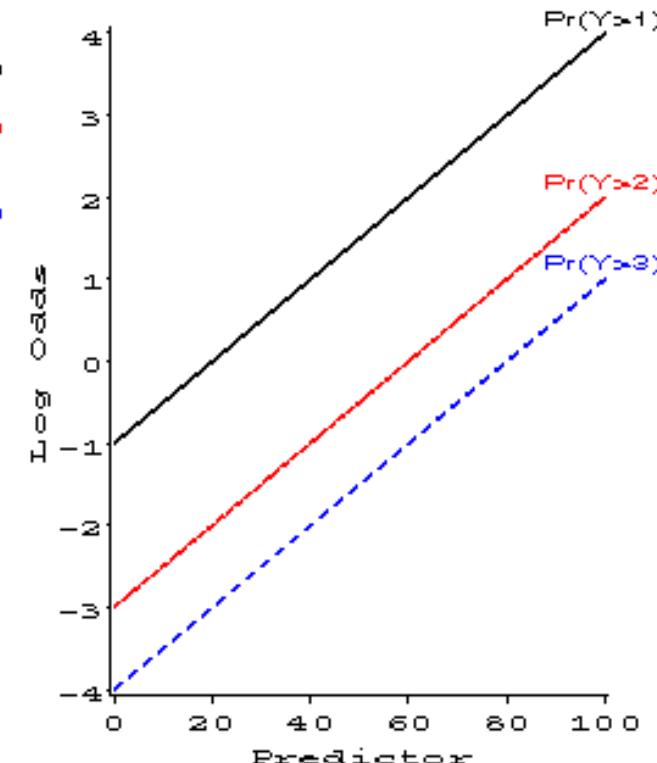
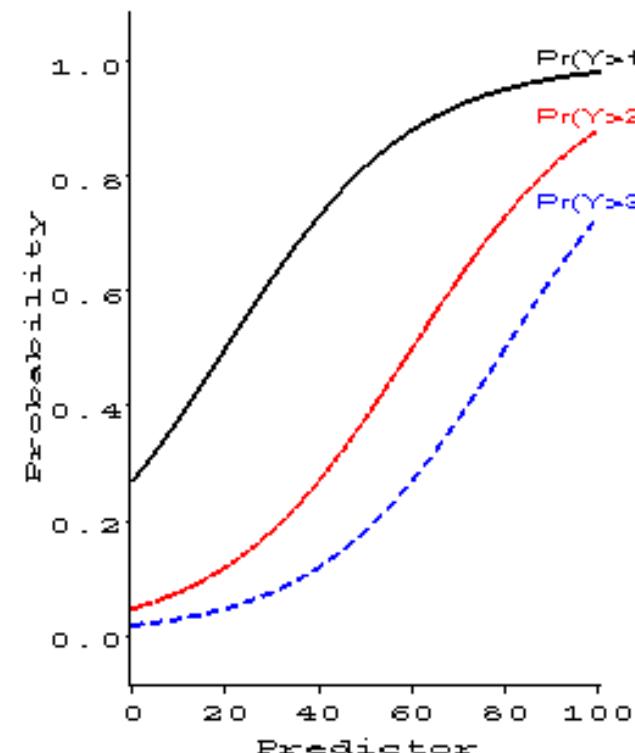
Proportional Odds Model

- A given predictor variable (x), which is the independent variable on the x-axis
- The probability (left panel) of a response being greater than ($P(Y>j)$) a certain value (j ; $j=1,2,3$) increases with continuous predictor values
 - Notice, with response values of $Y=1, 2, 3$, or 4 possible, there's only three curves, with three intercepts ($\alpha_{>1}, \alpha_{>2}, \alpha_{>3}$)
 - The 4th category is simply $1 - \alpha_{>1}$



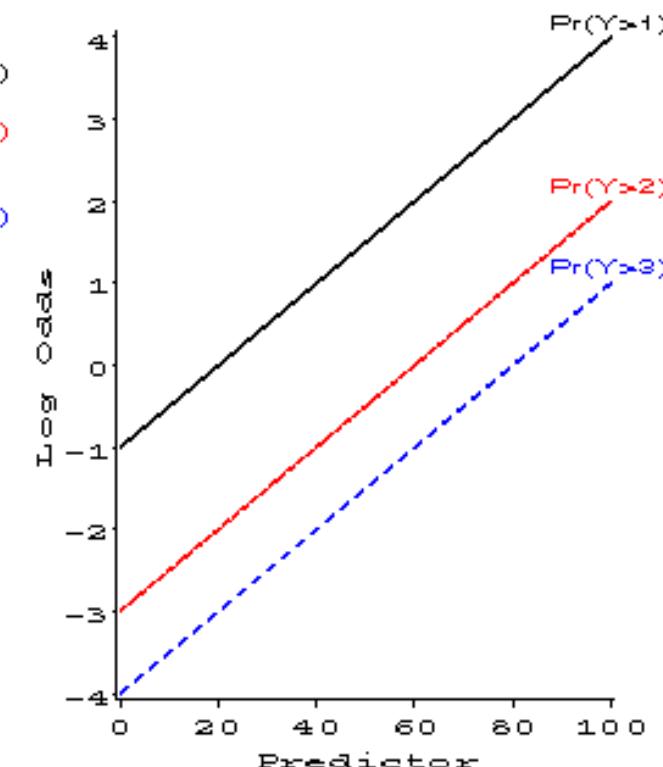
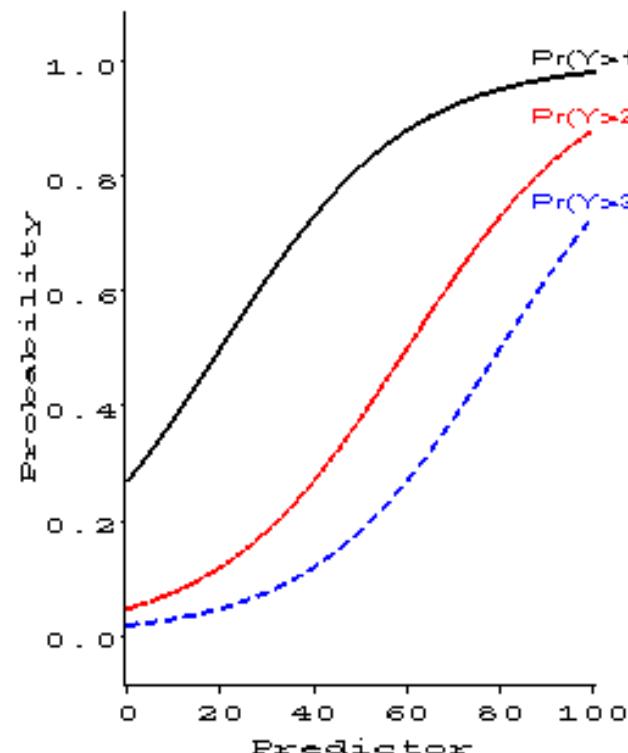
Proportional Odds Model

- With response values of $Y=1, 2, 3$, or 4 possible, could also determine probabilities of response being less than or equal to a value
 - $P(Y \leq 1), P(Y \leq 2), P(Y \leq 3), P(Y \leq 4)$
- Would still only produce three curves, with three intercepts ($\alpha_{\leq 1}, \alpha_{\leq 2}, \alpha_{\leq 3}$)
- The 4th category ($Y=4$) is simply $1 - \alpha_{\leq 3}$



Proportional Odds Model

- The right panel depicts the log odds (β) of each predictor variable
- This estimable parameter is the “slope”
- Notice how the intercepts (α) change with each response (Y) value, but the slope (β) is constant for each Y_j
- The slopes are parallel/proportional to each other over the range of observed predictor variable (x) values



Proportional Odds Model in R

- R uses a package called “MASS”
- MASS includes a function called polynomial regression
 - Notice not **binomial** (as in logistic regression), but **polynomial**, because we’re dealing with ordered categorical data (ordinal data)
 - Function in R: “`polr`”
- The “`polr`” function has option to run a variance-covariance matrix
 - Uses a “Hessian” matrix
 - Asks for “`Hess=`”: Can write “`FALSE`” to not use, “`TRUE`” to use

Proportional Odds Model in R

- There is no “base” model
- Unlike logistic regression, proportional odds models are not linear reparameterizations of a *baseline-category* model
- The intercept (α_j) is the log-odds of falling into or below ($P \leq j$) when any/all predictor variable (x) values are 0
 - The intercept parameter itself is a representation of a “base” model without covariates

FULL MODEL

```
#### Cumulative logit model (FULL MODEL)---  
install.packages("MASS")  
library(MASS)  
fit.pain1<-polr(painscore~Cmax+WT+GENDER, data=orldata, Hess=TRUE)  
summary(fit.pain1)
```

Call:

```
polr(formula = painscore ~ Cmax + WT + GENDER, data = orldata,  
      Hess = TRUE)
```

Coefficients: β

	Value	Std. Error	t value
Cmax	-0.017414	0.003930	-4.4311
WT	0.008617	0.007781	1.1075
GENDERMale	-0.218827	0.379210	-0.5771

Intercepts: α

	Value	Std. Error	t value
Complete pain relief Moderate pain relief	-1.3224	0.6424	-2.0584
Moderate pain relief No pain relief	-0.1781	0.6240	-0.2854
No pain relief slight pain relief	1.0846	0.6423	1.6887

Residual Deviance: 251.6615

AIC: 263.6615

Proportional Odds Model in R

BASE MODEL

```
fit.pain0<-polr(painscore~1, data=orldata, Hess=TRUE)  
summary(fit.pain0)
```

Call:
polr(formula = painscore ~ 1, data = orldata, Hess = TRUE)

No coefficients

Intercepts:

	Value	Std. Error	t value
Complete pain relief Moderate pain relief	-0.8001	0.2162	-3.7005
Moderate pain relief No pain relief	0.0800	0.2002	0.3999
No pain relief slight pain relief	1.1527	0.2341	4.9229

Residual Deviance: 275.1637
AIC: 281.1637

FULL MODEL

```
fit.pain1<-polr(painscore~Cmax+WT+GENDER, data=orldata, Hess=TRUE)  
summary(fit.pain1)
```

Call:
polr(formula = painscore ~ Cmax + WT + GENDER, data = orldata,
Hess = TRUE)

Coefficients:

	Value	Std. Error	t value
Cmax	-0.017414	0.003930	-4.4311
WT	0.008617	0.007781	1.1075
GENDERMale	-0.218827	0.379210	-0.5771

Intercepts:

	Value	Std. Error	t value
Complete pain relief Moderate pain relief	-1.3224	0.6424	-2.0584
Moderate pain relief No pain relief	-0.1781	0.6240	-0.2854
No pain relief slight pain relief	1.0846	0.6423	1.6887

Residual Deviance: 251.6615
AIC: 263.6615

Proportional Odds Model in R

- “Base model” is not very informative
- Probability of a response being less than or equal to one of the ordinal values, as opposed to greater than that value, is only being estimated by the intercept
- No real value in just that
- Need predictor variables (x) in model to analyze an exposure/response relationship

$$\text{logit}[P(Y \leq j)] = \alpha_j + \beta_1 x_1 + \beta_2 + \cdots + \beta_n x_n$$

$$P(Y \leq j) = \frac{\exp(\alpha_j + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_n x_n)}{1 + \exp(\alpha_j + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_n x_n)}$$

Proportional Odds Model in R - FULL

- Cmax was the only significant predictor variable in the full model
- Remove body weight and gender as covariates (reduced model)
- The only intercept that was significantly predictive of response was the alpha for the probability of achieving complete relief vs moderate
 - $P(Y \leq 3) / P(Y > 3)$

```
ctable1<-coef(summary(fit.pain1))
p1<-pnorm(abs(ctable1[,"t value"]),lower.tail=FALSE)*2
(ctable1<-cbind(ctable1,'p value' =p1))
ctable1
```

	Value	Std. Error	t value	p value
Cmax	-0.017413597	0.003929865	-4.4310933	9.375650e-06
WT	0.008617246	0.007780865	1.1074920	2.680813e-01
GENDERMale	-0.218827406	0.379210018	-0.5770612	5.638981e-01
Complete pain relief Moderate pain relief	-1.322362007	0.642433351	-2.0583645	3.955516e-02
Moderate pain relief No pain relief	-0.178057138	0.623975256	-0.2853593	7.753689e-01
No pain relief slight pain relief	1.084602271	0.642278184	1.6886799	9.128079e-02

Proportional Odds Model in R

- Reduced model has a lower AIC than Full model
 - Better model fit than Full mode
- Cmax still a significant predictor
- Each intercept is also now significantly predictive of probability of achieving a particular response (Y_j)
 - Was not the case in Full model

REDUCED MODEL

```
### Reduced model---
fit.pain_red<-polr(painscore~Cmax,data=orldata, Hess=TRUE)
summary(fit.pain_red)
summary(fit.pain1)
Call:
polr(formula = painscore ~ Cmax, data = orldata, Hess = TRUE)

Coefficients:
            Value Std. Error t value
Cmax -0.01721  0.003877 -4.439

Intercepts:
                                         Value Std. Error t value
Complete pain relief|Moderate pain relief -1.7872  0.3298 -5.4199
Moderate pain relief|No pain relief        -0.6541  0.2666 -2.4531
No pain relief|slight pain relief         0.5869  0.2671  2.1974

Residual Deviance: 253.0473
AIC: 261.0473
```

*AIC (full) = 263.66

	Value	Std. Error	t value	p value
Cmax	-0.0172086	0.003876619	-4.439075	9.034644e-06
Complete pain relief Moderate pain relief	-1.7872295	0.329753168	-5.419901	5.963209e-08
Moderate pain relief No pain relief	-0.6541220	0.266649880	-2.453112	1.416263e-02
No pain relief slight pain relief	0.5868930	0.267088409	2.197374	2.799378e-02

Different Representation of Ordinal Responses

- Change response orders from text to numeric
- Re-run the Reduced model while treating the ordered responses as numeric
- Will still result in 3 intercepts, 3 probabilities:
 1. $\log \frac{P(Y \leq 1)}{P(Y > 1)}$
Probability of no pain relief: $P(Y = 1) = P(Y \leq 1)$
 2. $\log \frac{P(Y \leq 2)}{P(Y > 2)}$
Probability of slight pain relief: $P(Y = 2) = P(Y \leq 2) - P(Y \leq 1)$
Probability of moderate pain relief: $P(Y = 3) = P(Y \leq 3) - P(Y \leq 2)$
 3. $\log \frac{P(Y \leq 3)}{P(Y > 3)}$
Probability of complete pain relief: $P(Y = 4) = 1 - P(Y \leq 3)$

Comparing Reduced Models

```
### Reduced model---  
fit.pain_red<-polr(painscore~Cmax,data=orldata, Hess=TRUE)  
summary(fit.pain_red)  
summary(fit.pain1)
```

Call:
polr(formula = painscore ~ Cmax, data = orldata, Hess = TRUE)

Coefficients:

	Value	Std. Error	t value
Cmax	-0.01721	0.003877	-4.439

Intercepts:

	Value	Std. Error	t value
Complete pain relief Moderate pain relief	-1.7872	0.3298	-5.4199
Moderate pain relief No pain relief	-0.6541	0.2666	-2.4531
No pain relief slight pain relief	0.5869	0.2671	2.1974

Residual Deviance: 253.0473

AIC: 261.0473

	Value	Std. Error	t value	p value
Cmax	-0.0172086	0.003876619	-4.439075	9.034644e-06
Complete pain relief Moderate pain relief	-1.7872295	0.329753168	-5.419901	5.963209e-08
Moderate pain relief No pain relief	-0.6541220	0.266649880	-2.453112	1.416263e-02
No pain relief slight pain relief	0.5868930	0.267088409	2.197374	2.799378e-02

```
### Different representation of categories--  
fit.pain_red2<-polr(factor(ps)~Cmax, data=orldata, Hess=TRUE)  
summary(fit.pain_red2)  
ctable_red2<-coef(summary(fit.pain_red2))  
p_red2<-pnorm(abs(ctable_red2[,"t value"]),lower.tail=FALSE)*2  
(ctable_red2<-cbind(ctable_red2,'p value' =p_red2))
```

Call:
polr(formula = factor(ps) ~ Cmax, data = orldata, Hess = TRUE)

Coefficients:

	Value	Std. Error	t value
Cmax	0.02277	0.004139	5.501

Intercepts:

	Value	Std. Error	t value
1 2	-0.3319	0.2796	-1.1872
2 3	1.0734	0.3055	3.5130
3 4	2.2352	0.3658	6.1096

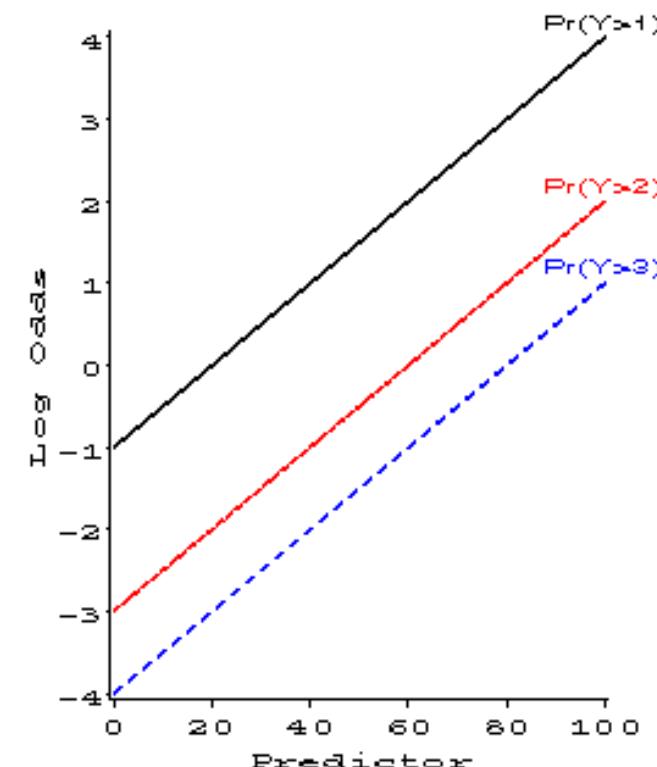
Residual Deviance: 238.7697

AIC: 246.7697

	Value	Std. Error	t value	p value
Cmax	0.02276659	0.00413887	5.500678	3.783326e-08
1 2	-0.33190318	0.27957875	-1.187155	2.351667e-01
2 3	1.07335569	0.30553921	3.512988	4.430969e-04
3 4	2.23519913	0.36584737	6.109649	9.985031e-10

Inferences of Parameters

- The intercepts (α) estimates are the log odds of having a particular response (Y_j) when any/all predictor variable (x) values are zero
- The slope (β) estimates are the log odds of having a particular response (Y_j) when that associated predictor variable (x) has a *specific* value
 - As the value of x changes, so too does the log odds of achieving that particular Y_j , but the probability is ***proportional*** to the change in the value of x



Confidence Intervals

- Because the intercepts (α) and slope(s) (β) are *estimates*, need to calculate the uncertainty associated with estimating
 - 95% CI

```
> exp(0.02276659)
[1] 1.023028
> exp(confint(fit.pain_red2))
Waiting for profiling to be done...
      2.5 %    97.5 %
1.015060 1.031715
```

```
> summary(orddata)
```

ID	painscore	ps	Cmax
Min. : 1.00	Complete pain relief:31	Min. :1.00	Min. : 0.00
1st Qu.: 25.75	Moderate pain relief:21	1st Qu.:2.00	1st Qu.: 0.00
Median : 50.50	No pain relief :24	Median :3.00	Median : 27.50
Mean : 50.50	slight pain relief :24	Mean :2.59	Mean : 50.25
3rd Qu.: 75.25		3rd Qu.:4.00	3rd Qu.: 98.00
Max. :100.00		Max. :4.00	Max. :200.00

TRT	isTRT	WT	GENDER
40mg CTM_IR:50	Min. :0.0	Min. : 30.00	Female:45
Placebo :50	1st Qu.:0.0	1st Qu.: 56.00	Male :55
	Median :0.5	Median : 65.50	
	Mean :0.5	Mean : 70.77	
	3rd Qu.:1.0	3rd Qu.: 82.50	
	Max. :1.0	Max. :138.00	

Confidence Intervals

- Because the intercepts (α) and slope(s) (β) are *estimates*, need to calculate the uncertainty associated with estimating
 - 95% CI

```
> exp(0.02276659)
[1] 1.023028
> exp(confint(fit.pain_red2))
Waiting for profiling to be done...
      2.5 %    97.5 %
1.015060 1.031715
```

- Interpreted as the odds of having complete pain relief ($Y = 4$), as compared to other categories of response ($Y = 1, 2$, or 3), increases by **1.023** per unit (ng/mL) increase in Cmax

$$\frac{P(Y > 3)}{P(Y \leq 3)}$$

Check Proportional Odds Assumption

- Hallmark of this model is that the slope estimate (β) is the same for all categories of responses (Y)
 - Results in parallel logits
- Need to test this assumption by plotting values of β when logistic regressions are performed set of binomial response variables
- “qlogis” is R code for finding the quantile of logistic regression
- Create a vector of binary responses:
 - “ $Y \geq 1$ ” = `qlogis(mean(y>=1))`
 - If $Y \geq 1$, then labeled a success ($Y=1$)
 - If $Y < 1$, then labeled a failure ($Y=0$)
 - Repeat for $Y=2, Y=3, Y=4$

Check Proportional Odds Assumption

- R automatically breaks the continuous predictor variable (C_{max}) into tertiles (thirds)

- In lowest tertile ($C_{max}=0$), i.e. the intercept-only model, β estimates are NOT the same for $Y=3$ and $Y=4$
 - Should be...
 - For upper two tertiles, β estimates are the same for $Y=3$ and $Y=4$

```
##### Check proportional odds assumption-----
install.packages("Hmisc")
library(Hmisc)
sf <- function(y) {
  c(`Y>=1` = qlogis(mean(y >= 1)),
    `Y>=2` = qlogis(mean(y >= 2)),
    `Y>=3` = qlogis(mean(y >= 3)),
    `Y>=4` = qlogis(mean(y >= 4)))}
(s <- with(orddata, summary(as.numeric(painscore) ~ Cmax, fun = sf)))

as.numeric(painscore)      N= 100
+-----+-----+-----+-----+
|           |N   |Y>=1|Y>=2       |Y>=3       |Y>=4   |
+-----+-----+-----+-----+
|Cmax     | 0    | 50 |Inf  | 1.8152900| 0.94446161|-0.94446161|
|          | [ 55,100)| 26 |Inf  | 0.6359888|-0.81093022|-1.2039728|
|          | [100,200] | 24 |Inf  |-0.5108256|-1.60943791|-1.60943791|
+-----+-----+-----+-----+
|Overall|           |100|Inf  | 0.8001193|-0.08004271|-1.1526795|
+-----+-----+-----+-----+
```

Check Proportional Odds Assumption

- Normalize the first set of variable coefficients (β estimates for $Y=2$) to be zero
- Model wasn't significantly predictable for $Y=1$ to $Y=2$, only $Y=2$ to $Y=3$, and $Y=3$ to $Y=4$

```
#### Normalize the first set of coefficients to be zero
s[, 5] <- s[, 5] - s[, 3]
s[, 4] <- s[, 4] - s[, 3]
s[, 3] <- s[, 3] - s[, 3]
# print
s
```

```
as.numeric(painscore)      N= 100

+-----+-----+-----+-----+
|           |N   |Y>=1|Y>=2|Y>=3       |Y>=4       |
+-----+-----+-----+-----+
|Cmax    | 0    | 50|Inf |0    |-0.8708284|-2.759752|
|        |[ 55,100)| 26|Inf |0    |-1.4469190|-1.839962|
|        |[100,200] | 24|Inf |0    |-1.0986123|-1.098612|
+-----+-----+-----+-----+
|Overall|          |100|Inf |0    |-0.8801620|-1.952799|
+-----+-----+-----+-----+
```

Check Proportional Odds Assumption

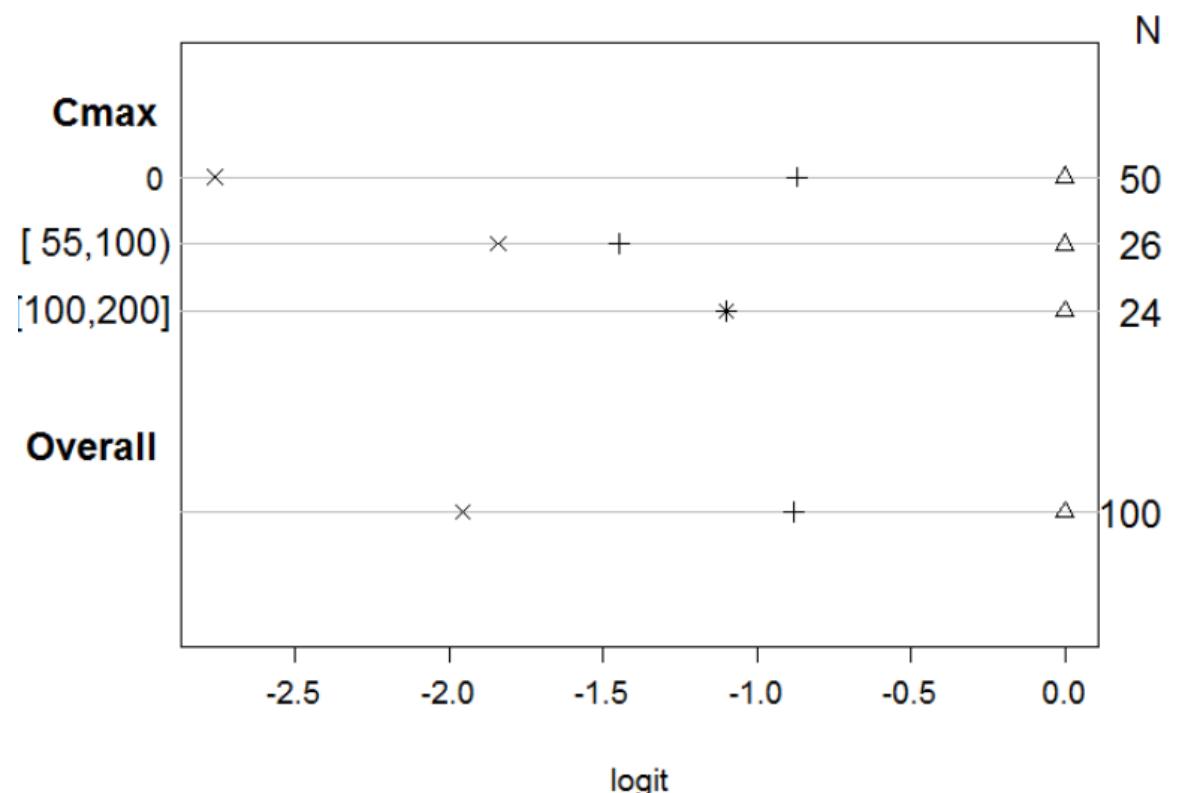
- Normalize the first set of variable coefficients (β estimates for $Y=2$) to be zero
- Model wasn't significantly predictable for $Y=1$ to $Y=2$, only $Y=2$ to $Y=3$, and $Y=3$ to $Y=4$

```
#### Normalize the first set of coefficients to be zero
s[, 5] <- s[, 5] - s[, 3]
s[, 4] <- s[, 4] - s[, 3]
s[, 3] <- s[, 3] - s[, 3]
# print
s

##### Plot to check proportional odds assumption-----
plot(s, which = 1:4, pch = 1:4, xlab = "logit", main = " ",
      xlim = range(s[,3:5]), cex=1.2)

as.numeric(painscore)      N= 100
```

	N	Y>=1 Y>=2 Y>=3 Y>=4
Cmax	0	50 Inf 0 -0.8708284 -2.759752
	[55,100)	26 Inf 0 -1.4469190 -1.839962
	[100,200]	24 Inf 0 -1.0986123 -1.098612
Overall		100 Inf 0 -0.8801620 -1.952799



Plotting the Predicted Probabilities

- What is the purpose of analyzing the exposure (Cmax)- response (pain relief) relationship?
 - So we can get an idea of how much drug exposure is needed to induce the optimal drug response
 - Knowing the exposure needed, can determine the dose amount and frequency
- Can use the proportional odds model to predict probability of obtaining each level of response (pain relief)
 - If only have empirical, observed exposure data on a limited exposure range, building the model can then be used to simulate what probabilities of pain relief would be on exposure values not observed
 - Extrapolated data

Plotting the Predicted Probabilities

$$\text{logit}[P(Y \leq j)] = \alpha_i + \beta_1 C_{\max}$$

$$\text{logit}[P(Y \leq 1)] = \alpha_1 + 0.0228(C_{\max})$$

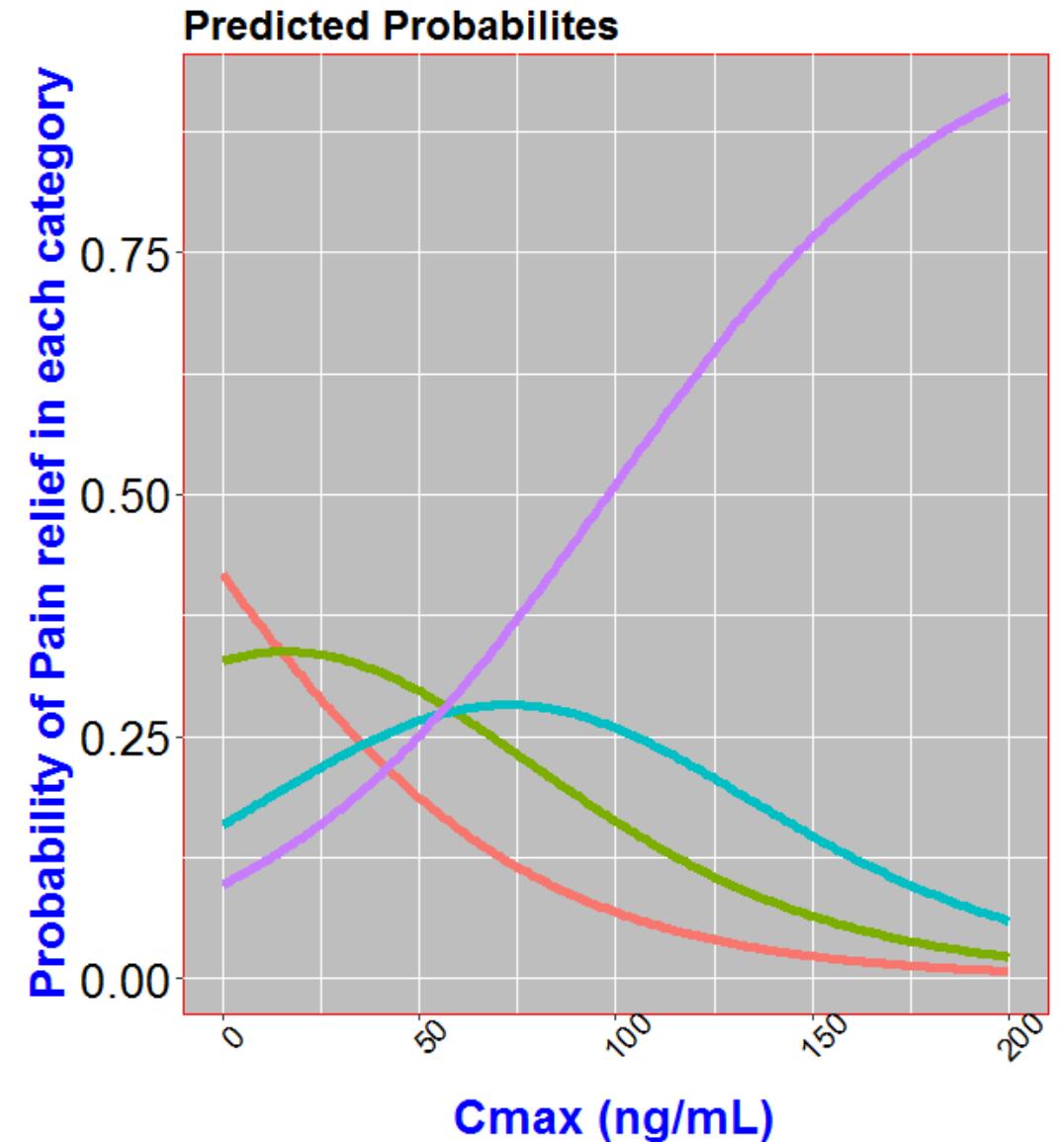
$$\text{logit}[P(Y \leq 2)] = \alpha_2 + 0.0228(C_{\max})$$

$$\text{logit}[P(Y \leq 3)] = \alpha_3 + 0.0228(C_{\max})$$

$$\text{logit}[P(Y \leq 4)] = 1 - \text{logit}[P(Y \leq 3)]$$

As increase drug exposure (C_{max}):

- Probability of complete pain relief (Y=4; purple) increases (Odds Ratio increases 1.02 from 0 – 98ng/mL ng/mL)
- Probability of no pain relief Y=1, red) decreases
- Probability of slight pain relief (Y=2; green) is the same for a range, then decreases
- Probability of moderate pain relief (Y=3), increases for a range of exposures, then tails off as complete pain relief is more likely



```
#### Predictions for the responses-----
newdat<- data.frame(Cmax = seq(0,200,length.out = 200 ))
newdat<- cbind(newdat, predict(fit.pain_red2, newdat, type = "probs"))
## show first few rows
head(newdat)

##### Rearrange new dataset-----
library(reshape2)
lnewdat<-melt(newdat, id.vars = c("Cmax"), variable.name = "Level", value.name = "Probability")

#### Plot of probabilities-----
plot3<-ggplot(lnewdat, aes(x = Cmax, y = Probability, colour = Level)) +
  geom_line(size=2) +
  xlab("Cmax (ng/mL)") +
  ylab("Probability of Pain relief in each category")
plot3+ theme(panel.background = element_rect(fill='grey', colour='red'))+
  theme(axis.title.y = element_text(colour = 'blue', size = 20, face='bold'))+
  theme(axis.text.y = element_text(size = 20, colour='black'))+
  theme(axis.title.x = element_text(colour = 'blue', size = 20, face='bold'))+
  theme(axis.text.x = element_text(size = 14, colour='black', angle=45))+  
  ggtile("Predicted Probabilités")+
  theme(plot.title = element_text(lineheight=.8, face="bold",size = 18))+  
  theme(legend.position="right")+
  theme(legend.title=element_blank())+
  theme(legend.text = element_text(colour="black", size = 10))
```

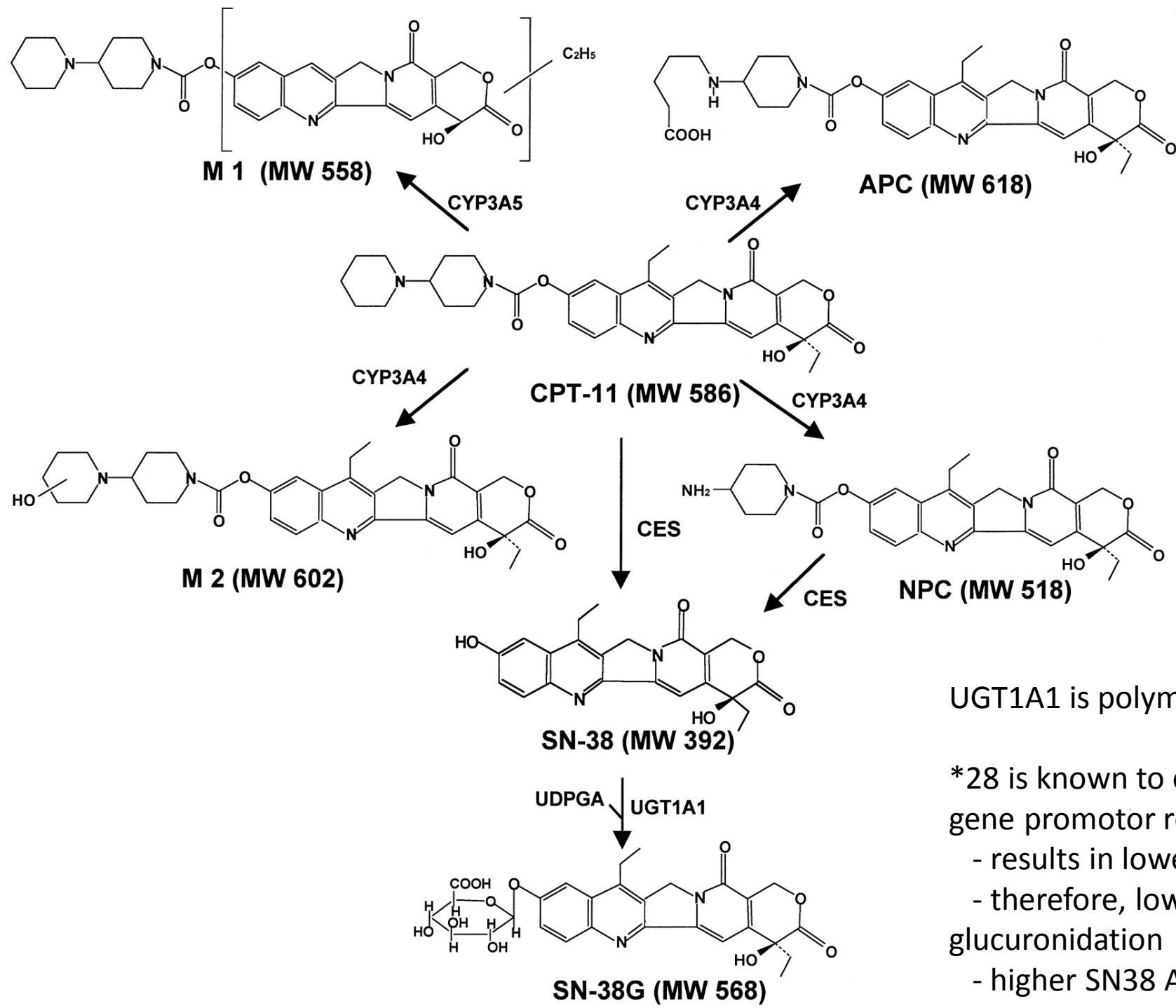
AM Break

Published Example

- Use of a proportional odds model to assess the relationship between a targeted anti-cancer agent (irinotecan) and its dose-limiting side effect, diarrhea
- First, a little about the drug itself

Irinotecan

- A topoisomerase I inhibitor that inhibits a cell's ability to repair DNA damage
- Works best when given with DNA-damaging chemotherapy
 - Induction of DNA damage by chemotherapeutics will require topoisomerases to repair DNA
 - With irinotecan blocking DNA repair via topoisomerase inhibition, induces apoptosis (cell death)
- Received FDA approval in 1996 for colorectal cancers with use of 5-fluorouracil + leucovorin



UGT1A1 is polymorphic.

*28 is known to encode errors in gene promotor region

- results in lower gene expression
- therefore, lower SN38 glucuronidation
- higher SN38 AUC

Irinotecan

- FDA outlined dose reductions based on severity of side effects
- These are all *retrospective* dose reductions
 - Patients have to experience this to then be adjusted
- What if we could know the exposure of the entity causing this tox, and tailor the dose to achieve an effective, yet sub-toxic dose?

Table 2. Recommended Dose Modifications for CAMPTOSAR/5-Fluorouracil (5-FU)/Leucovorin (LV) Combination Schedules

Patients should return to pre-treatment bowel function without requiring antidiarrhea medications for at least 24 hours before the next chemotherapy administration. A new cycle of therapy should not begin until the granulocyte count has recovered to $\geq 1500/\text{mm}^3$, and the platelet count has recovered to $\geq 100,000/\text{mm}^3$, and treatment-related diarrhea is fully resolved. Treatment should be delayed 1 to 2 weeks to allow for recovery from treatment-related toxicities. If the patient has not recovered after a 2-week delay, consideration should be given to discontinuing therapy.

Toxicity NCI CTC Grade ^a (Value)	During a Cycle of Therapy	At the Start of Subsequent Cycles of Therapy ^b
No toxicity	Maintain dose level	Maintain dose level
Neutropenia 1 (1500 to $1999/\text{mm}^3$) 2 (1000 to $1499/\text{mm}^3$) 3 (500 to $999/\text{mm}^3$) 4 ($<500/\text{mm}^3$)	Maintain dose level \downarrow 1 dose level Omit dose until resolved to \leq grade 2, then \downarrow 1 dose level Omit dose until resolved to \leq grade 2, then \downarrow 2 dose levels	Maintain dose level Maintain dose level \downarrow 1 dose level \downarrow 2 dose levels
Neutropenic fever	Omit dose until resolved, then \downarrow 2 dose levels	
Other hematologic toxicities	Dose modifications for leukopenia or thrombocytopenia during a cycle of therapy and at the start of subsequent cycles of therapy are also based on NCI toxicity criteria and are the same as recommended for neutropenia above.	
Diarrhea 1 (2-3 stools/day $>$ pretx ^c) 2 (4-6 stools/day $>$ pretx) 3 (7-9 stools/day $>$ pretx) 4 (≥ 10 stools/day $>$ pretx)	Delay dose until resolved to baseline, then give same dose Omit dose until resolved to baseline, then \downarrow 1 dose level Omit dose until resolved to baseline, then \downarrow 1 dose level Omit dose until resolved to baseline, then \downarrow 2 dose levels	Maintain dose level Maintain dose level \downarrow 1 dose level \downarrow 2 dose levels
Other nonhematologic toxicities^d 1 2 3 4	Maintain dose level Omit dose until resolved to \leq grade 1, then \downarrow 1 dose level Omit dose until resolved to \leq grade 2, then \downarrow 1 dose level Omit dose until resolved to \leq grade 2, then \downarrow 2 dose levels <i>For mucositis/stomatitis decrease only 5-FU, not CAMPTOSAR.</i>	Maintain dose level Maintain dose level \downarrow 1 dose level \downarrow 2 dose levels <i>For mucositis/stomatitis decrease only 5-FU, not CAMPTOSAR.</i>

^a National Cancer Institute Common Toxicity Criteria (version 1.0)

^b Relative to the starting dose used in the previous cycle

^c Pretreatment

^d Excludes alopecia, anorexia, asthenia

Clinical Pharmacology and Therapeutics

- In 2002, a Dutch group published results of their study of irinotecan pharmacokinetics and toxicity in 109 patients with any type of solid tumor after at least one dose per patient
 - Doses ranged from 100 – 350 mg/m²
- 44 patients (40.3%) received a second dose
 - Indicative how toxic irinotecan is that 60% couldn't tolerate a second dose
 - Diarrhea was most common nonhematologic toxicity
 - Leukopenia and neutropenia were most common, most severe toxicities resulting in dose-reductions
 - Some cases, diarrhea severe, non-responsive to anti-diarrhea treatments
- Previous studies noted how diarrhea was correlated with exposures of drug (parent irinotecan and/or metabolites)
 - No one had ever “modeled” this exposure/response relationship to predict the probability of achieving a particular ordered grade (severity) of diarrhea based on how much drug that person was exposed to

Methods

- Collected PK samples up to 60 hrs post dose
- Measured plasma concentrations of irinotecan, SN38, and SN38-G
- Exposure metrics: AUC_{0-60hr} for all 3 compounds
- Response metric: grade of diarrhea
 - Grade 0: 0-1 stools more than normal during pretreatment
 - Grade 1: 2-3 stools more than normal during pretreatment
 - Grade 2: 4-6 stools more than normal during pretreatment
 - Grade 3: 7-9 stools more than normal during pretreatment
 - Grade 4: 10+ stools more than normal during pretreatment

Methods

- Proportional Odds Model
- A single predictor variable (x):

AUC_{0-60hr}=

- One slope parameter (β)
- 4 intercept parameters (α)
 - 5 ordered grades (0-4), so 4 intercepts
 - $P(Y=5) = 1 - (P(Y \leq 4)$

- Each drug's AUC modeled separately

regression, similar as used for dry mouth score.²⁴ If $Y_n = (Y_0, Y_1, \dots)$ is the vector of the diarrhea scores for individual patients, the probability for Y_n larger than or equal to the score m ($m = 0, 1, 2, 3$, and 4) can be expressed as follows:

$$g[P(Y_n \geq m)] = \text{logit}(p) = \text{Int} + f_d$$

in which

$$\text{logit}(p) = \log[p/(1 - p)]$$

and

$$p = \exp(\text{Int} + f_d)/(1 + \exp[\text{Int} + f_d])$$

$$\text{Int} = \sum_{L=0}^m \mu_{Li}$$

$$f_d = \text{Slop} \cdot \text{AUC}(0-60)$$

where g is the logit function of a probability (p is short notation for $P[Y_n \geq m]$) and $\mu_{Li} \geq 0$ specifies the baseline probabilities of diarrhea scores and are

Results

- To induce a significant change, need a drop in objective function value (OFV) of at least 3.84 pts to be significant at $p<0.05$
- AUC_{0-60hr} for parent irinotecan and SN38-G significantly predicted grade of diarrhea
 - Not SN38 AUC_{0-60hr}
- Many PK samples couldn't measure SN38-G, only n=51 patients had AUC data
 - Reduced data set

Table IV. Difference in objective function value (ΔOFV)

Drug	ΔOFV	
	Full data set (n = 104)	Reduced data set (n = 51)
Irinotecan	-4.29*	-3.514
SN-38	-2.468	-1.672
SN-38G	—	-8.43†
Biliary index	—	-0.522

ΔOFV , Change in objective function value.

* $P = .04$.

† $P = .0038$.

Results

- Below are the parameters for the intercepts and slope within each drug model

Table V. Summary of population pharmacodynamic estimates

Parameter	Irinotecan (<i>n</i> = 104)		SN-38 (<i>n</i> = 104)		SN-38G (<i>n</i> = 51)	
	Estimate	SE	Estimate	SE	Estimate	SE
Intercept, m = 1	-0.55	0.42	-0.31	0.35	-0.62	0.37
Intercept, m = 2	-1.23	0.19	-1.22	0.19	-1.10	0.26
Intercept, m = 3	-1.69	0.31	-1.68	0.31	-1.58	0.39
Intercept, m = 4	-1.16	0.49	-1.15	0.49	-0.90	0.53
Slope*	4.26E-05	2.0E-05	1.20E-03	7.9E-04	1.87E-04	6.0E-05

m, Diarrhea score; E, scientific expression of number.

*Slope for linear drug effect on the logit scale.

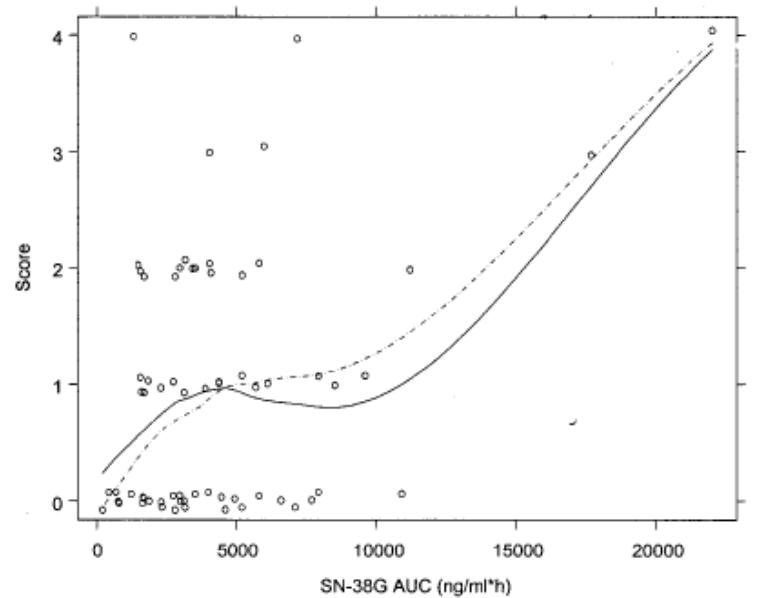
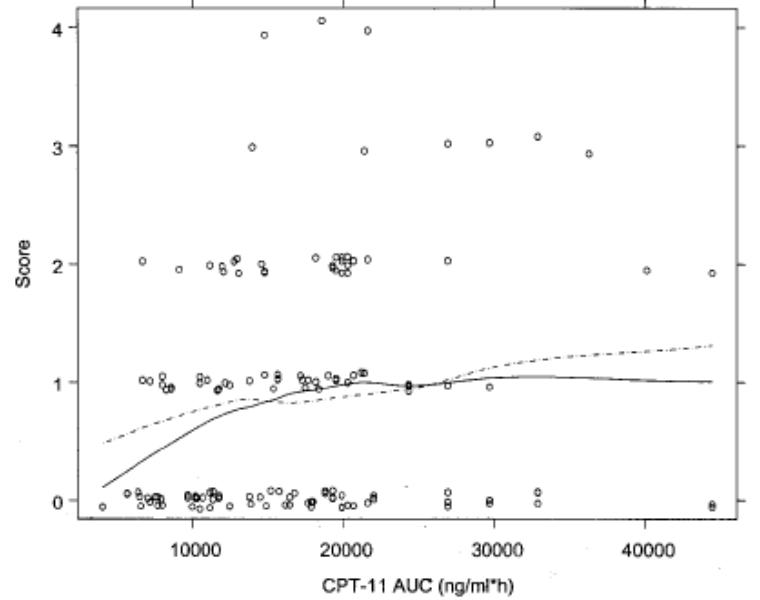


Fig 3. Diarrhea scores versus the area under the concentration–time curve from time zero to 60 hours [AUC(0–60)] for irinotecan (CPT-11; *top panel*) and SN-38G (*bottom panel*). The *solid lines* are the smoothed lines for observed scores; the *broken lines* are the smoothed lines for the predicted scores, which are simulated from the final pharmacokinetic-pharmacodynamic models.

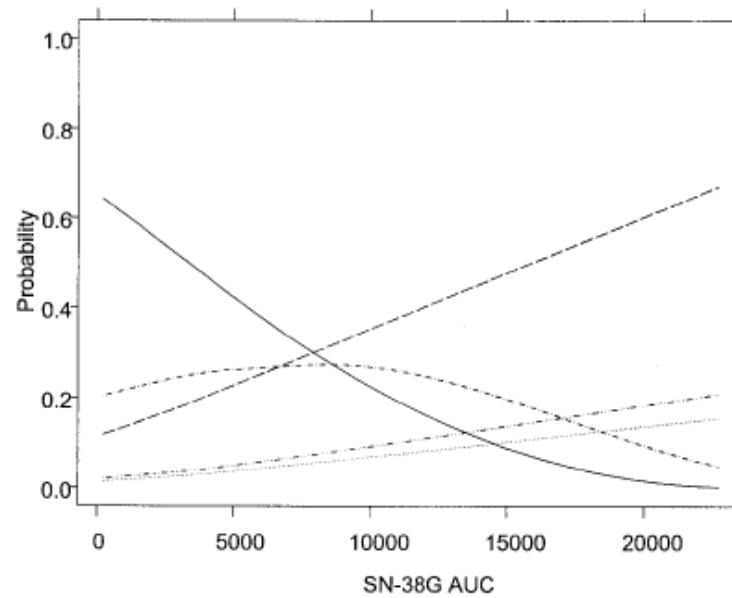
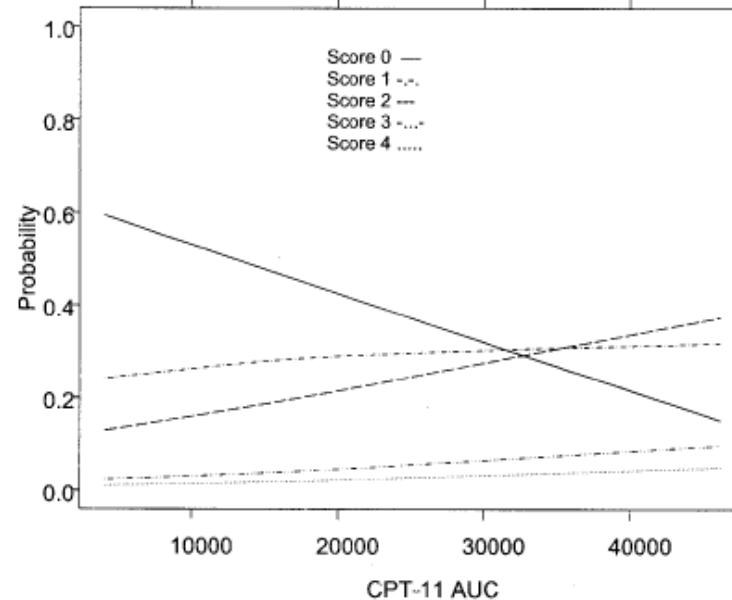


Fig 4. Predicted probabilities for diarrhea score 0 (*solid lines*), diarrhea score 1 (*dashed-dotted lines*), diarrhea score 2 (*dashed lines*), diarrhea score 3 (*lines with sequence of dash, 3 dots, dash, 3 dots*), and diarrhea score 4 (*dotted lines*) versus irinotecan AUC(0–60) (in nanograms per milliliter per hour; *top panel*) and SN-38G AUC(0–60) (in nanograms per milliliter per hour; *bottom panel*). CPT-11, Irinotecan.

Impact

- It is known that patients carrying at least one (of two) copies of *UGT1A1*28* are susceptible to reduced conversion of SN38 to SN38-G
- FDA has not officially recommended any dose reductions based on genotype, but models can predict extent of decreased metabolism
 - Therefore, can personalize dose based on genotype
- PK models can identify doses to achieve a certain exposure of irinotecan, SN38, and SN38-G
- NOW, can tailor the dose to achieve an exposure that will not induce Grade 2+ diarrhea (i.e. severe)

Lunch

Exposure/Response Modeling III-B

Continuous or Discrete Variables vs Count (Discrete) Responses
Using General Estimation Equations (GEE) and
Generalized Linear Mixed Effects Models (GLMM)

Types of Clinical Data Variables

- **Qualitative or Categorical** – a variable with categories
 - Ordinal – meaningful ranking or scale, but not quantified
 - Example: pain status (mild, moderate, severe)
 - Example: cancer staging or grades (1, 2, 3, or 4)
 - Nominal – NO meaningful rankings
 - Binomial – yes/no; gender (male/female), genotype (for some genes)
 - Other – blood types, race, genotype (for some genes)
- **Quantitative** – a variable with numeric values
 - Continuous – measured on a continuous scale
 - Example: age, body weight, drug concentration
 - Discrete – measured on a discrete scale
 - Example: number of seizures within a time period (count data), survival (event data)

Additional Types of Count Data

Count Data

- Number of days in a hospital
- Number of lesions observed

Rate Data (Longitudinal)

- Number of seizures in a week
- Number of doctor visits in a month

Case Study: Seizures

- Randomized, placebo-controlled trial of the anti-seizure Drug A for treatment of seizures
- 150 subjects
 - Randomized (1:2, placebo:Drug A)
- Demographics: age at baseline, gender
- Exposure metric: presence of drug vs placebo
- Response metric: seizure count over 3 months

Research Question

- Is the drug effective in reducing the number (rate) of seizures (over 3 months) vs placebo?
- Are there any confounding covariates?
 - Age, Gender

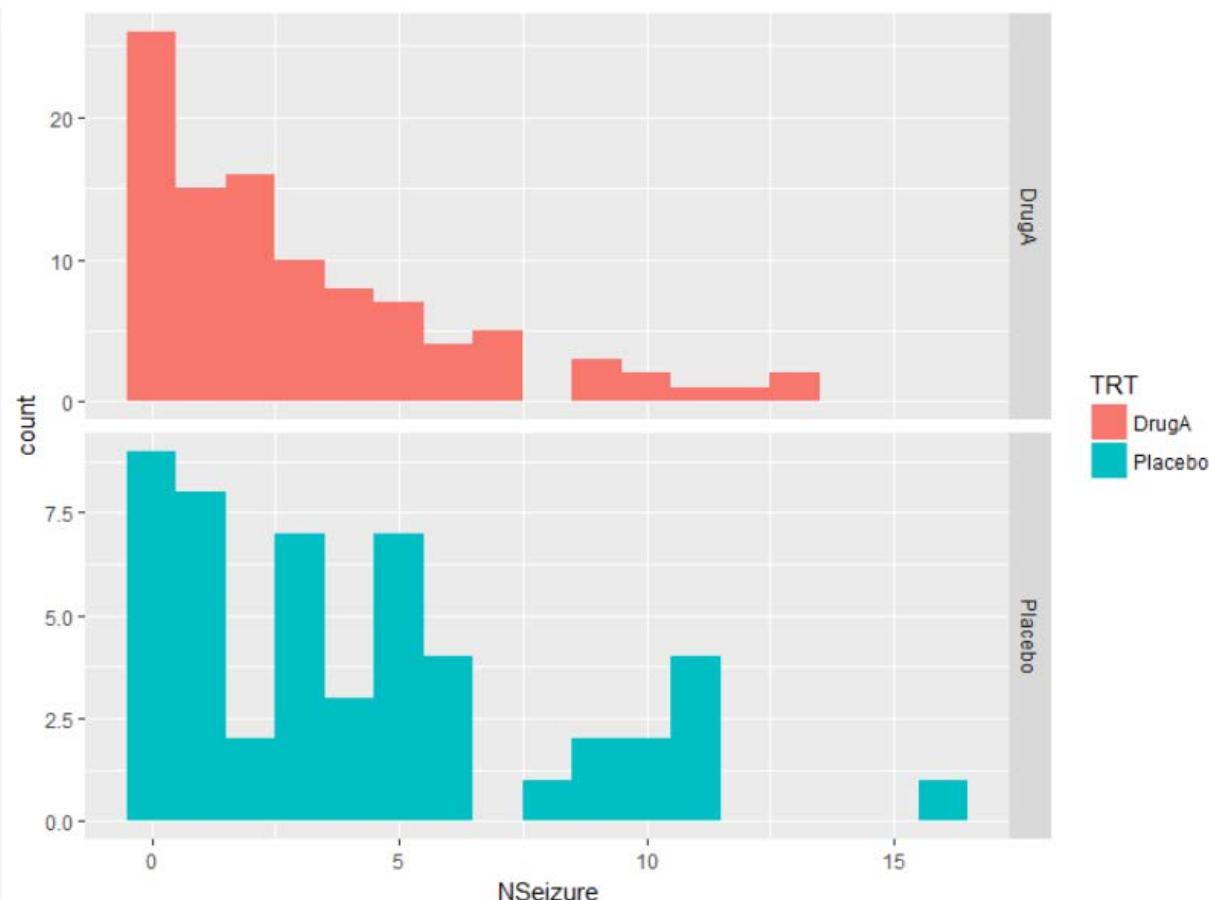
Seizure Data Exploration

```
seiz.data<- read.csv("countdata.csv", sep=",")  
head(seiz.data)
```

ID	Age	Gender	isF	TRT	NSeizure
1	43	Female	1	Placebo	6
2	55	Male	0	Placebo	2
3	76	Female	1	Placebo	5
4	93	Male	0	Placebo	0
5	95	Female	1	Placebo	2
6	35	Male	0	Placebo	4

```
> summary(seiz.data)  
ID           Age       Gender      isF        TRT      NSeizure  
Min. : 1.00  Min. : 18.00  Female:81  Min. :0.00  DrugA :100  Min. : 0.000  
1st Qu.: 38.25 1st Qu.: 38.00  Male :69   1st Qu.:0.00  Placebo:50  1st Qu.: 1.000  
Median : 75.50  Median : 45.00          Median :1.00          Median : 2.000  
Mean   : 75.50  Mean   : 48.47          Mean   :0.54          Mean   : 3.393  
3rd Qu.:112.75 3rd Qu.: 56.75          3rd Qu.:1.00          3rd Qu.: 5.000  
Max.  :150.00  Max.  :112.00          Max.  :1.00          Max.  :16.000  
> save image("C:/Users/neerc/Desktop/NCT Projects/Classes/FAFS/RinTech84_ExnResultsinR March201
```

```
##### Plot of Seizure data by ####treatment-----  
library(ggplot2)  
plot<-ggplot(seiz.data, aes(NSeizure, fill = TRT)) +  
  geom_histogram(binwidth = 1) +  
  facet_grid(TRT ~ .,scales = "free")  
plot
```



Seizure Data Exploration

```
> summary(seiz.data)
   ID          Age       Gender      isF      TRT      NSeizure
Min. : 1.00  Min. :18.00  Female:81  Min. :0.00  DrugA :100  Min. : 0.000
1st Qu.: 38.25 1st Qu.:38.00  Male :69   1st Qu.:0.00  Placebo: 50  1st Qu.: 1.000
Median : 75.50  Median :45.00
Mean   : 75.50  Mean   :48.47
3rd Qu.:112.75 3rd Qu.:56.75
Max.  :150.00  Max.  :112.00

```

NSeizure
Min. : 0.000
1st Qu.: 1.000
Median : 2.000
Mean : 3.393
3rd Qu.: 5.000
Max. :16.000

```
> ### use tapply to get table of mean, sd for #seizures over 3 months---
> with(seiz.data, tapply(NSeizure, TRT, function(x) {sprintf("M (SD) = %1.2f (%1.2f)", 
+   mean(x), sd(x)) }))
                    DrugA           Placebo
"M (SD) = 2.99 (3.20)" "M (SD) = 4.20 (3.86)"
> |
```

It appears there's fewer seizures over 3 month period in patients given Drug A vs placebo

Count Data Analysis

- These distributions are used to model discrete events that occur infrequently
- Can also be used to describe the number of occurrences of an event over a given time interval
 - e.g., number of seizures in a month, number of cancer occurrences in a year
- Normally associated with “count” data
- Essentially working with numerous instances of binary outcomes
 - Can perform multiple logistic regressions, one at each time point
 - Impractical

Generalized Linear Model Framework

- A way to model the **mean response** against predictor variable(s) through a logistic (or a logit) function
 - No “residual” error like in linear regression for continuous variables (obs-pred)
 - In logistic regression, error: $\text{Var}(\varepsilon_i) = P_i * (1-P_i)$
 - Error is a function of probability of having success or no success
- Generalized Linear Model (GLM)
 - 3 components:
 1. **Distributional assumption** (random component)
 - Identifies response (Y) and its probability distribution
 2. **Systematic component**
 - Specifies predictor variables used in a linear predictor function
$$g[E(Y)] = \beta_0 + \beta_1 x_1 + \beta_2 x_2$$
 3. **Link function**
 - Specifies $g(\cdot)$, links the random (1) and systematic (2) components together

Generalized Linear Model – Logistic Regression

1) Distributional Assumption for Response (Y)

- Exponential family

$$f(Y : \mu) = \exp[a(Y)*b(\mu) + c(\mu) + d(Y)]$$

μ : parameter (not population mean)

b: natural parameter of the distribution

- Binomial distribution: a member of the exponential family

$$f(Y : p) = \binom{n}{Y} p^Y (1 - p)^{n-Y}$$

$$b(\mu) = \log\left(\frac{\mu}{1 - \mu}\right) = \log\left(\frac{p}{1 - p}\right)$$

*If $n=1$, Y follows a Bernoulli distribution

- for pos/neg responses

b: natural parameter of *binomial* distribution

Count Data uses Poisson Distribution Assumption

- Poisson distribution used to model:
 - Discrete data that occurs rarely
 - ✓ • Number of occurrences of an event over a given time interval

Poisson Distributional Assumptions for Count Data Response (Y)

- Probability of an event ($P(Y=1)$) is proportional (\sim) to the length of the time interval
 - $P(\text{cancer in time } t) = \lambda * t$
 - Event rate (λ) is **constant** throughout time
 - Events occur **independently**
- If these assumptions are met, then distribution of number (count) of events in time interval (t) can be considered to follow a Poisson distribution
 - Mean, $\mu = \lambda * t$

Poisson Distribution

- The probability mass function of a Poisson response variable (Y) is:

$$P(Y = y) = \frac{e^{-\lambda} * \lambda^y}{y!}, y = 0, 1, 2, 3, \dots$$

- Example: suppose the incidence of a certain type of disease in a given region is 250 cases per year
- What is the *probability* that there will be exactly 135 cases within the next 6 months?

- t=0.5

$$P(135|\lambda = 125) = \frac{e^{-125} * 125^{135}}{135!} = 0.0232$$

- There's a 2.32% chance of observing 135 cancers in next 6 months, knowing that the *average* rate is 125 cancers in 6 months (250/yr)

Poisson Distribution

- The probability mass function of a Poisson response variable (Y) is:

$$P(Y = y) = \frac{e^{-\lambda} * \lambda^y}{y!}, y = 0, 1, 2, 3, \dots$$

$E(Y) = \lambda$ ** expected number of counts/events

$RATE = \lambda/t$ ** expected number of events/time period

$\text{Var}(Y) = \lambda$ ** variance around the expected mean

- Fundamental feature of Poisson distribution is that mean and variance of a parameter are the **SAME**
 - When counts (# events) are higher, the mean is higher, and so too the variability around that mean

Generalized Linear Model

2) Systematic Component

- Relates effect of predictor variables (covariates) to the *transformed mean response* through a linear model

$$g[E(Y)] = \beta_0 + \beta_1 x_1 + \beta_2 x_2$$

3) Link Function

- Links the random component (distributional assumption) to the systematic component
- Modeling count data requires a ***log-link*** function

$$\log(\lambda) = \beta_0 + \beta_1 x_1$$

* **Poisson regression model**

$$\lambda = \exp(\beta_0 + \beta_1 x_1)$$

Log-Link Function for Modeling Event Rates

- If the time of observation or number of subjects in a group changes, need an “offset” term

$$\log(\lambda/t) = \beta_0 + \beta_1 x_1$$

$$\log(\lambda) - \log(t) = \beta_0 + \beta_1 x_1$$

$$\log(\lambda) = \log(t) + \beta_0 + \beta_1 x_1$$

Log(t) is a known offset term

Maximum Likelihood Estimation (MLE)

$l(\beta)$ = likelihood estimate of parameter

MLE is the value (estimate) that reduces the objective function value (OFV), aka
-2*log likelihood

Log-likelihood function:

MLE cannot be solved; estimates determined by an iterative process

$$l(\beta) = \sum_{i=1}^n \log(P(Y = y)) = \sum_{i=1}^n \log\left(\frac{\exp(\lambda_i) * \lambda_i^{y_i}}{y_i!}\right)$$

$$= \sum_{i=1}^n [y_i x'_i \beta - \exp(x'_i \beta) - \log(y_i !)]$$

Take the first derivative w.r.t. the parameter, and equate to 0

$$l'(\beta) = \sum_{i=1}^n [y_i - \exp(x'_i \hat{\beta})] x_j = 0$$

β is the parameter being estimated by MLE

Poisson Regression – BASE Model

- Analyze seizure data to see if drug associated with the number (count) of seizure events over a 3-month period

```
fit.pois.base<- glm(NSeizure ~ 1, family = "poisson", data = seiz.data)
summary(fit.pois.base)

Call:
glm(formula = NSeizure ~ 1, family = "poisson", data = seiz.data)

Deviance Residuals:
    Min      1Q  Median      3Q     Max 
-2.6051 -1.5307 -0.8198  0.8142  4.9408 

Coefficients:
            Estimate Std. Error z value Pr(>|z|)    
(Intercept) 1.22181   0.04432  27.57   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for poisson family taken to be 1)

Null deviance: 534.43  on 149  degrees of freedom
Residual deviance: 534.43  on 149  degrees of freedom
AIC: 893.17

Number of Fisher Scoring iterations: 5
```

Poisson Regression – FULL Model

```
fit.pois.full<-glm(NSeizure ~ relevel(TRT,"Placebo") + Gender+Age,  
                    family = "poisson", data = seiz.data)  
summary(fit.pois.full)
```

All predictor variables are significant

```
Call:  
glm(formula = NSeizure ~ relevel(TRT, "Placebo") + Gender + Age,  
     family = "poisson", data = seiz.data)
```

Deviance Residuals:

Min	1Q	Median	3Q	Max
-3.7385	-1.7733	-0.3596	0.7923	4.7521

Coefficients:

	Estimate	Std. Error	z value	Pr(> z)
(Intercept)	2.377007	0.157354	15.106	< 2e-16 ***
relevel(TRT, "Placebo")DrugA	-0.357134	0.090066	-3.965	7.33e-05 ***
GenderMale	-0.226751	0.090442	-2.507	0.0122 *
Age	-0.018032	0.003096	-5.824	5.75e-09 ***

Signif. codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for poisson family taken to be 1)

Null deviance: 534.43 on 149 degrees of freedom

Residual deviance: 475.40 on 146 degrees of freedom

AIC: 840.14

Releveled treatment factor is estimating the effect of Drug A *relative to Placebo*

Gender factor is estimating the effect of males *relative to females*

AIC lower than Base Model

Number of Fisher Scoring iterations: 5

Poisson Regression – FULL Model

- Full model: $\log(\lambda) = \beta_0 + \beta_1(\text{DrugA}) + \beta_2(\text{Gender}) + \beta_3(\text{Age})$
 $\log(\lambda) = 2.38 - 0.357(\text{DrugA}) - 0.227(\text{Male}) - 0.018(\text{Age})$
- How well does this model describe the data?
- AKA, how good is the model fit?

```
### Assess goodness of fit--  
with(fit.pois, cbind(res.deviance = deviance, df = df.residual,  
    pvalue = pchisq(deviance, df.residual, lower.tail = FALSE)))  
  
Number of Fisher Scoring iterations: 5  
  
> with(fit.pois, cbind(res.deviance = deviance, df = df.residual,  
+     pvalue = pchisq(deviance, df.residual, lower.tail = FALSE)))  
    res.deviance   df      pvalue  
[1,]    475.4021 146 1.623943e-36  
>
```

Poisson Regression Model

- One of the main tenants of this model was that the mean and variance of a response were the **SAME**
- To verify this is the case, need to assess equality of mean, variance

```
> with(seiz.data, tapply(NSeizure, TRT, function(x) {sprintf("M (SD) = %1.2f (%1.2f)",  
+   mean(x), sd(x)) }))  
          DrugA           Placebo  
"M (SD) = 2.99 (3.20)" "M (SD) = 4.20 (3.86)"
```

- Variance = $(SD)^2$
- $\text{Var}_{\text{DrugA}}$ vs $\text{Mean}_{\text{DrugA}}$: 10.24 vs 2.99 (Var > Mean)
- $\text{Var}_{\text{Placebo}}$ vs $\text{Mean}_{\text{Placebo}}$: 14.89 vs 3.86 (Var > Mean)
- Overdispersion

Overdispersion

How does overdispersion impact the model inference?

- Parameter estimates (β) **WILL NOT** change
- SE around β **WILL** change
 - Higher SE will lead to inflated type I error

How do we account for overdispersion?

1. i) Estimate, then ii) adjust, the dispersion parameter (2-stage approach)

$$\text{Var}(Y) = \phi * \lambda \quad \text{If } \phi = 1, \text{ then no overdispersion}$$

2. Quasi-likelihood approach (1-stage approach)

- Dispersion factor is modeled

- ✓ 3. Negative binomial regression (most common approach)

Negative Binomial Distribution

- Probability mass function:

$$P(Y = y|k, \lambda) = \frac{\Gamma(y + k)}{\Gamma(k)\Gamma(y + 1)} \left(\frac{k}{\lambda + k}\right)^k \left(1 - \frac{k}{\lambda + k}\right)^y$$

where y (response) = 0, 1, 2, ...

Γ is similar to factorial

λ is estimate for mean, variance

k is involved in dispersion factor

$$E(Y) = \lambda$$

$$\text{Var}(Y) = \lambda + \lambda^2/k$$

($1/k$ is the dispersion factor)

Negative Binomial Regression – FULL Model

```
##### Negative binomial regression -----
library(MASS)
fit.nb <- glm.nb(NSeizure ~ relevel(TRT,"Placebo") + Gender + Age, data = seiz.data)
summary(fit.nb)

Call:
glm.nb(formula = NSeizure ~ relevel(TRT, "Placebo") + Gender +
    Age, data = seiz.data, init.theta = 1.32321106, link = log)

Deviance Residuals:
    Min      1Q  Median      3Q     Max 
-2.2214 -1.0752 -0.1920  0.4147  2.0436 

Coefficients:
            Estimate Std. Error z value Pr(>|z|)    
(Intercept)  2.44308   0.30269   8.071 6.95e-16 ***
relevel(TRT, "Placebo")DrugA -0.36471   0.17743  -2.055 0.039833 *  
GenderMale   -0.21872   0.17079  -1.281 0.200320    
Age          -0.01945   0.00541  -3.595 0.000325 ***  
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for Negative Binomial(1.3232) family taken to be 1)

Null deviance: 189.98 on 149 degrees of freedom
Residual deviance: 172.05 on 146 degrees of freedom
AIC: 699.92

Number of Fisher Scoring iterations: 1

Theta:  1.323
Std. Err.: 0.245
2 x log-likelihood: -689.920
```

Releveled treatment factor is estimating the effect of Drug A *relative to Placebo*

*Now, gender isn't a significant predictor variable.

AIC much lower than FULL Model using Poisson distribution

Theta = k, so 1/theta is the dispersion factor/parameter
(1/k)

Negative Binomial Regression – REDUCED Model

```
### Remove gender, since nonsignificant, thus final model--  
fit.nb1<-glm.nb(NSeizure ~ relevel(TRT,"Placebo")+Age, data = seiz.data)  
summary(fit.nb1)
```

```
call:  
glm.nb(formula = NSeizure ~ relevel(TRT, "Placebo") + Age, data = seiz.data,  
       init.theta = 1.295627343, link = log)
```

```
Deviance Residuals:  
    Min      1Q  Median      3Q     Max  
-2.1652 -1.0994 -0.2277  0.4433  2.1678
```

```
Coefficients:
```

	Estimate	Std. Error	z value	Pr(> z)
(Intercept)	2.355590	0.295083	7.983	1.43e-15 ***
relevel(TRT, "Placebo")DrugA	-0.377009	0.178408	-2.113	0.034585 *
Age	-0.019425	0.005435	-3.574	0.000351 ***

```
---
```

```
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
(Dispersion parameter for Negative Binomial(1.2956) family taken to be 1)
```

```
Null deviance: 187.80  on 149  degrees of freedom  
Residual deviance: 171.73  on 147  degrees of freedom  
AIC: 699.54
```

```
Number of Fisher Scoring iterations: 1
```

Releveled treatment factor is estimating the effect of Drug A relative to Placebo

Theta: 1.296
Std. Err.: 0.237

2 x log-likelihood: -691.540

AIC slightly lower than FULL Model

Poisson Regression – FINAL Model

- Final model: $\log(\lambda) = \beta_0 + \beta_1(\text{DrugA}) + \beta_2(\text{Age})$
 $\log(\lambda) = 2.36 - 0.377(\text{DrugA}) - 0.019(\text{Age})$
- How well does this model describe the data?
- AKA, how good is the model fit?

```
##### Goodness of fit of the final model -----
with(fit.nb1, cbind(res.deviance = deviance, df = df.residual,
  p = pchisq(deviance, df.residual, lower.tail = FALSE)))
### p-value for chi-sq test, testing deviance between final and saturated, is nonsignificant---
## meaning we accept null hypothesis that current model adequately fits data--
```

```
> with(fit.nb1, cbind(res.deviance = deviance, df = df.residual,
+   p = pchisq(deviance, df.residual, lower.tail = FALSE)))
  res.deviance df          p
[1]    171.7295 147 0.07973327
```

Current model not different than Ideal model (deviance), p=0.08.
Means current model is optimal

Poisson Regression – FINAL Model

$$\log(\lambda) = \beta_0 + \beta_1(\text{DrugA}) + \beta_2(\text{Age})$$

$$\log(\lambda) = 2.36 - 0.377(\text{DrugA}) - 0.019(\text{Age})$$

```
##### Confidence interval of the parameter estimates -----
(est <- cbind(Estimate = coef(fit.nb1), confint(fit.nb1)))
```

```
> (est <- cbind(Estimate = coef(fit.nb1), confint(fit.nb1)))
Waiting for profiling to be done...
      Estimate      2.5 %      97.5 %
(Intercept)  2.35558989  1.76915854  2.956564637
relevel(TRT, "Placebo")DrugA -0.37700926 -0.73202277 -0.029392425
Age          -0.01942506 -0.03026872 -0.008628924
```

Poisson Regression – FINAL Model

$$\log(\lambda) = \beta_0 + \beta_1(\text{DrugA}) + \beta_2(\text{Age})$$

$$\log(\lambda) = 2.36 - 0.377(\text{DrugA}) - 0.019(\text{Age})$$

```
##### Confidence interval of the parameter estimates -----
(est <- cbind(Estimate = coef(fit.nb1), confint(fit.nb1)))
##### Exponentiate the estimates-----
exp(est)
```

```
> (est <- cbind(Estimate = coef(fit.nb1), confint(fit.nb1)))
Waiting for profiling to be done...
              Estimate      2.5 %      97.5 %
(Intercept)    2.35558989  1.76915854  2.956564637
relevel(TRT, "Placebo")DrugA -0.37700926 -0.73202277 -0.029392425
Age            -0.01942506 -0.03026872 -0.008628924
```

```
> exp(est)
              Estimate      2.5 %      97.5 %
(Intercept)    10.5443470  5.8659153 19.2317900
relevel(TRT, "Placebo")DrugA  0.6859097  0.4809352  0.9710353
Age            0.9807624  0.9701848  0.9914082
```

Interpretation of Final Model

$$\log(\lambda) = 2.36 - 0.377(\text{DrugA}) - 0.019(\text{Age})$$

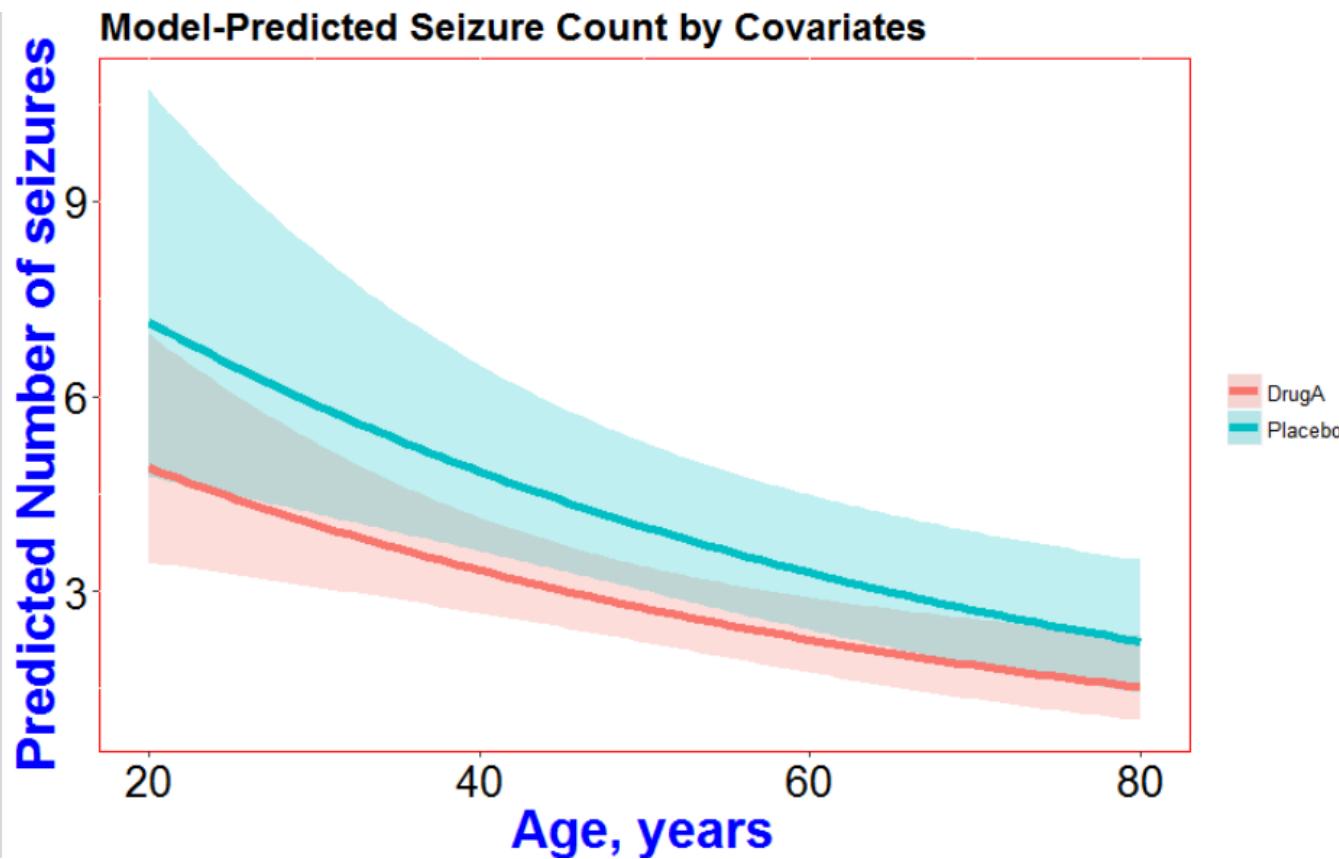
```
> exp(est)
              Estimate      2.5 %      97.5 %
(Intercept) 10.5443470 5.8659153 19.2317900
relevel(TRT, "Placebo")DrugA  0.6859097 0.4809352  0.9710353
Age          0.9807624 0.9701848  0.9914082
```

- When comparing number of seizures (count) in patients taking Drug A (vs placebo), parameter estimate is 0.686
 - Inference: Drug A reduces seizures by 32.4% ($1-0.686$) compared to placebo
- Age is a continuous covariate (OR based on unit change (yr) in age)
 - Parameter estimate is 0.98
 - Inference: There is a 2% ($1-0.98$) reduction in seizures q3mo with each year increase in age

Plotting the Model-Predicted Seizure Counts

$$\log(\lambda) = 2.36 - 0.377(\text{DrugA}) - 0.019(\text{Age})$$

- Can plot the mean estimate (λ) for *predicted* seizure count by age and treatment (drug A vs placebo)
 - Estimate +/- 95% confidence interval
 - As get older, seizure count predicted to drop (2% reduction in number of seizures that occur every 3 months, with each passing year of age)
 - Concurrently, those people taking Drug A have 36% fewer seizures vs people taking placebo



1) Make a new dataframe
(newdata2)

- simulating ages from 20-80 yr, n=100
- simulating treatment assignment (1:2 randomization of placebo:drugA, n=100)

```
##### Predictions-----  
newdata2 <- data.frame(Age = rep(seq(from = 20, to = 80,  
length.out = 100), 2),  
TRT = factor(rep(1:2, each = 100), levels = 1:2, labels =  
levels(seiz.data$TRT)))
```

1) Make a new dataframe
(newdata2)

- simulating ages from 20-80 yr, n=100
- simulating treatment assignment (1:2 randomization of placebo:drugA, n=100)

```
##### Predictions-----
newdata2 <- data.frame(Age = rep(seq(from = 20, to = 80,
                                         length.out = 100), 2),
                        TRT = factor(rep(1:2, each = 100), levels = 1:2, labels =
                                         levels(seiz.data$TRT)))
newdata2 <- cbind(newdata2, predict(fit.nb1, newdata2,
                                         type = "Link", se.fit=TRUE))
```

2) Bind results of final model to the new dataframe

- provides model algorithm and parameter estimates, with SE

1) Make a new dataframe
(newdata2)

- simulating ages from 20-80 yr, n=100
- simulating treatment assignment (1:2 randomization of placebo:drugA, n=100)

```
##### Predictions-----
newdata2 <- data.frame(Age = rep(seq(from = 20, to = 80,
                                         length.out = 100), 2),
                        TRT = factor(rep(1:2, each = 100), levels = 1:2, labels =
                                         levels(seiz.data$TRT)))
newdata2 <- cbind(newdata2, predict(fit.nb1, newdata2,
                                         type = "link", se.fit=TRUE))
newdata2 <- within(newdata2, {
  NumberofSeizures <- exp(fit)
  LL <- exp(fit - 1.96 * se.fit)
  UL <- exp(fit + 1.96 * se.fit)
})
```

2) Bind results of final model to the new dataframe

- provides model algorithm and parameter estimates, with SE

3) Within dataframe, calculate the lower 95% CI (LL) and upper limit (UL) using SE from model

4) Make a plot of
dataframe using ggplot2.

- layer of age with

#seizures

- “ribbons” are shaded
areas, represented here
by LL and UL based on
treatment status

- solid lines are the
parameter estimates
based on treatment
status

```
##### Predictions-----
newdata2 <- data.frame(Age = rep(seq(from = 20, to = 80,
                                         length.out = 100), 2),
                        TRT = factor(rep(1:2, each = 100), levels = 1:2, labels =
                                         levels(seiz.data$TRT)))
newdata2 <- cbind(newdata2, predict(fit.nb1, newdata2,
                                         type = "link", se.fit=TRUE))
newdata2 <- within(newdata2, {
  NumberofSeizures <- exp(fit)
  LL <- exp(fit - 1.96 * se.fit)
  UL <- exp(fit + 1.96 * se.fit)
})
plot1<-ggplot(newdata2, aes(Age, NumberofSeizures)) +
  geom_ribbon(aes(ymin = LL, ymax = UL, fill = TRT), alpha = .25) +
  geom_line(aes(colour = TRT), size = 2) +
  labs(x = "Age, years", y = "Predicted Number of seizures")
```

4) Make a plot of dataframe using ggplot2.

- layer of age with

```
#seizures
```

- “ribbons” are shaded areas, represented here by LL and UL based on treatment status

- solid lines are the parameter estimates based on treatment status

5) Add ggplot themes for aesthetics

```
##### Predictions-----
newdata2 <- data.frame(Age = rep(seq(from = 20, to = 80,
                                     length.out = 100), 2),
                       TRT = factor(rep(1:2, each = 100), levels = 1:2, labels =
                                     levels(seiz.data$TRT)))
newdata2 <- cbind(newdata2, predict(fit.nb1, newdata2,
                                    type = "link", se.fit=TRUE))
newdata2 <- within(newdata2, {
  NumberofSeizures <- exp(fit)
  LL <- exp(fit - 1.96 * se.fit)
  UL <- exp(fit + 1.96 * se.fit)
})
plot1<-ggplot(newdata2, aes(Age, NumberofSeizures)) +
  geom_ribbon(aes(ymin = LL, ymax = UL, fill = TRT), alpha = .25) +
  geom_line(aes(colour = TRT), size = 2) +
  labs(x = "Age, years", y = "Predicted Number of seizures")
plot1+ theme(panel.background = element_rect(fill='white', colour='red'))+
  theme(axis.title.y = element_text(colour = 'blue', size = 25, face='bold'))+
  theme(axis.text.y = element_text(size = 20, colour='black'))+
  theme(axis.title.x = element_text(colour = 'blue', size = 25, face='bold'))+
  theme(axis.text.x = element_text(size = 20, colour='black'))+
  ggtitle("Model-Predicted Seizure Count by Covariates")+
  theme(plot.title = element_text(lineheight=.8, face="bold",size = 18))+
  theme(legend.position="right")+
  theme(legend.title=element_blank())+
  theme(legend.text = element_text(colour="black", size = 10))
```

Longitudinal Data

- Longitudinal data is that which is collected over time from same patient
 - Monitoring a patient's body weight throughout study (patient has multiple body weight observations over time, likely changes, could be due to therapy)
 - Measuring a drug's plasma concentration (PK)
 - Have multiple drug concentration measurements per patient over time

Longitudinal Data

- Previously, we dealt with **single** observations per subject
 - Single time point
- Used models to see if certain variables (covariates) were significant predictors of the observed response
 - Binary responses (Bernoulli distribution)
 - With binary predictor variable – used Chi Squared Tests, Odds Ratio, Relative Risk
 - With 2+ predictor variables (discrete or continuous) – used logistic regression
 - Ordinal responses (cumulative probability) - used proportional odds models
 - Count responses (Poisson distribution, corrected for overdispersion by negative binomial distribution)
 - Poisson regression models

How to Model Longitudinal Data

- Have multiple responses (Y) for each subject
 - Y could be discrete or continuous
 - When Y is continuous, can perform linear or non-linear mixed effects modeling (*will not discuss that here*)
 - When Y is discrete, cannot simply use generalized linear models (GLM) that logistic regression and poisson regression models utilized
 - GLM ignores correlations between subjects
 - Mixed effects modeling (using for cont vs cont) accounts for this using random effects
- Need to use a model that accounts for inter-subject correlations with ***discrete*** longitudinal responses

Case Study

- Longitudinal discrete responses could be binary, ordinal, or count
- Use example of a respiratory drug
 - Binary responses observed on months 1,2, 3, and 4:
 - Status good = 1
 - Status poor = 0
 - Predictor variables (covariates):
 - Age, gender, baseline status (month 0), hospital, treatment group

Dataset Exploration

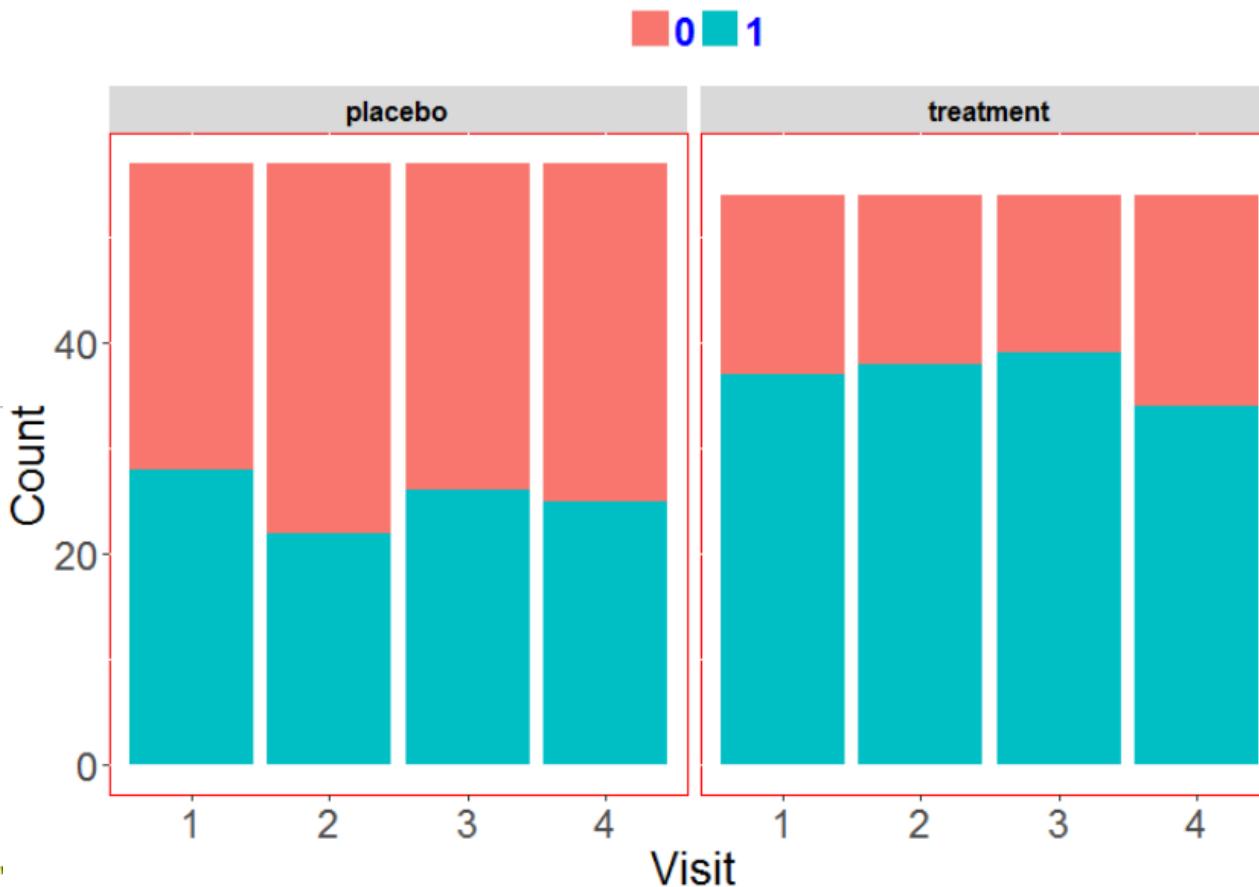
- Read in dataset
- Examine dataset header
- Assess data via tables
 - “addmargins” provides sum totals for both columns and rows
- Number of poor status (0) increased over time
- Number of good status (1) decreased over time
- Difficult to assess much from these

```
resp<-read.csv("resp.csv", sep=",")  
head(resp)  
  
  . . .  
  centre treatment sex age status month subject baseline nstat  
1     1 placebo female 46   poor    1      1    1   poor   0  
2     1 placebo female 46   poor    2      1    1   poor   0  
3     1 placebo female 46   poor    3      1    1   poor   0  
4     1 placebo female 46   poor    4      1    1   poor   0  
5     1 placebo female 28   poor    1      2    2   poor   0  
6     1 placebo female 28   poor    2      2    2   poor   0  
  
> ftable(addmargins(xtabs(~nstat+month,data=resp)))  
          month  1   2   3   4 Sum  
nstat  
0           46  51  46  52 195  
1           65  60  65  59 249  
Sum         111 111 111 111 444  
> ftable((xtabs(~treatment+nstat+month,data=resp)))  
          month  1   2   3   4  
treatment nstat  
placebo   0           29 35 31 32  
          1           28 22 26 25  
treatment 0           17 16 15 20  
          1           37 38 39 34  
> ftable((xtabs(~baseline+nstat+month,data=resp)))  
          month  1   2   3   4  
baseline nstat  
good     0           7 14  9 12  
          1           43 36 41 38  
poor    0           39 37 37 40  
          1           22 24 24 21  
.
```

Plot Data

```
##### Re arrange data for stacked bar chart-----
install.packages("reshape2")
library(reshape2)
count1<-ftable(xtabs(~treatment+nstat+month,data=resp))
### melt fnc converts data from a wide format to a
mydf = melt(count1)
names(mydf) = c("TRT","Outcome","Visit","Count")

###plot it--
##### Stacked bar chart-----
plot1<-ggplot(mydf, aes(x=Visit, y=Count, fill=Outcome )) +
  geom_bar(stat="identity")+
  facet_grid(.~TRT)
plot1+ theme(panel.background = element_rect(fill='white', colour='black'))+
  theme(axis.title.y = element_text(colour = 'black', size = 20))+ 
  theme(axis.text.y = element_text(size = 18))+ 
  theme(axis.title.x = element_text(colour = 'black', size = 20))+ 
  theme(axis.text.x = element_text(size = 18))+ 
  theme(plot.title = element_text(lineheight=.8, face="bold"))+
  theme(legend.position="top")+
  theme(legend.title=element_blank())+
  theme(legend.text = element_text(colour="blue", size = 18, face = "bold"))+
  theme(strip.text.x = element_text(size = 12, face = "bold",colour = "black"))
```



More “Good” statuses in DrugA vs placebo

Longitudinal Discrete Data Analysis

- Variance of longitudinal continuous outcomes is *independent* of means
 - Linear regression
- Variance of discrete outcomes at single time point is **dependent** on means
 - Based on type of discrete response, have Bernoulli and Poisson distributions
- Variance of longitudinal discrete outcomes is therefore more complicated
 - Have to factor in variance being a function of mean
 - Result: the correlation of responses between and within patients is a function of the mean
- Two analytical methods possible:
 1. Generalized estimating equations (GEE)
 2. Generalized linear mixed effects modeling (GLMM)

1. Generalized Estimating Equations (GEE)

- GEE is a method of estimation for marginal models
 - GEE is an approach to estimating parameters, not actually a model
- A marginal model is a ***population-averaged*** model
 - Model for mean response depends **only** on the covariates of interest and **not** on any random (i.e. inter-subject) effects
 - Mean response and models for association between observations are separately defined
 - No distributional assumptions for responses
 - Only a regression model for the mean response
 - Similar to logistic regression, except adding a component for correlation of responses within each individual

1. Generalized Estimating Equations (GEE)

$$g(\mu_{ij}) = x'_{ij}\beta$$

x' refers to matrix

- An extension of generalized linear model (GLM; only useful for single time point discrete responses)

3 parts to this marginal (population averaged) model:

1. Link function

- Mean of each response depends on the covariates

$$\log\left(\frac{\mu_{ij}}{1 - \mu_{ij}}\right) = x'_{ij}\beta$$

2. Variance $Var(Y_{ij}) = \phi v(\mu_{ij})$

- Depends on the mean response, given the effect of covariates
- ϕ is a scaling parameter needed for count data to control for overdispersion (not needed/used when response is binary)
- In this case study, response IS binary, so $\phi = 1$

$$Var(Y_{ij}) = \mu_{ij}(1 - \mu_{ij})$$

1. Generalized Estimating Equations (GEE)

$$g(\mu_{ij}) = x'_{ij}\beta$$

3. Within-subject association:

- Among the vector of repeated responses is assumed to be a function of a set of association parameters (α)

$$\log(OR) = \alpha$$

$$OR = \frac{\frac{P(Y_{ij} = 1|x_{ij} = 1)}{P(Y_{ij} = 0|x_{ij} = 1)}}{\frac{P(Y_{ij} = 1|x_{ij} = 0)}{P(Y_{ij} = 0|x_{ij} = 0)}}$$

Special Features of the Marginal Model

- Mean and variance of the data are modeled separately
- Regression parameters (β) have a population-averaged interpretation
- Parameter estimation in the marginal model is via GEE
 - GEE is an alternative to maximum likelihood estimation (MLE)

How does GEE work?

1. Perform a naïve linear regression, assuming observations within subjects are independent
2. Calculate the residuals from naïve model (observed – predicted), and estimate a working correlation matrix
3. Refit the regression coefficients for the correlation
 - An iterative process
4. Treat the within-subject correlation structure as a covariate

Correlation (Variance-Covariance) Matrices

1. Independent

- Diagonals are the variance ($\text{variance} = \text{sd}^2$) terms
- 3 time points (in this example), so a 3×3 structure
- Covariance (off-diagonal terms) = 0

$$\begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}$$

2. Exchangeable (most common)

- Assumes same covariance among all observations within subject for all subjects
- Covariance are the rho terms (ρ)

$$\begin{bmatrix} 1 & \rho & \rho \\ \rho & 1 & \rho \\ \rho & \rho & 1 \end{bmatrix}$$

3. Auto-Regressive (AR1)

- Correlations decrease (regress) among observations at each time point within subject

$$\begin{bmatrix} 1 & \rho & \rho^2 & \rho^3 \\ \rho & 1 & \rho & \rho^2 \\ \rho^2 & \rho & 1 & \rho \\ \rho^3 & \rho^2 & \rho & 1 \end{bmatrix}$$

Correlation Matrices

4. M-dependent

- Only correlates some time points within a subject, not all

$$\begin{bmatrix} 1 & \rho_1 & \rho_2 & 0 \\ \rho_1 & 1 & \rho_1 & \rho_2 \\ \rho_2 & \rho_1 & 1 & \rho_1 \\ 0 & \rho_2 & \rho_1 & 1 \end{bmatrix}$$

5. Unstructured

- A working correlation matrix
- Has no structure; could take any format, based on data
- Because of this, have to estimate a larger number of parameters for each subject
- If stack all estimated parameters, get a block diagonal matrix

$$\begin{bmatrix} 1 & \rho_1 & \rho_2 & \rho_3 \\ \rho_1 & 1 & \rho_5 & \rho_4 \\ \rho_2 & \rho_5 & 1 & \rho_6 \\ \rho_3 & \rho_4 & \rho_6 & 1 \end{bmatrix}$$

GEE in R

- Fit a marginal model to the “resp” dataset (FULL model)
 - Base model not very informative
- Designated correlation matrix structure (“corstr”) as *independent*
- Response is binary (poor (0) or good (1))
 - Family = binomial
- SE has two values
 - Naïve SE: model-based SE assuming correlation matrix structure is correct
 - Robust SE (aka Sandwich SE): calculated through empirically-corrected variance-covariance estimates, assuming the presence of heteroscedasticity or clustering
 - Uses Hessian (2nd partial derivative) square matrix from GEE solution

```
resp_gee1<-gee(nstat ~ centre + treatment + sex + baseline + age,  
                 data = resp, family = "binomial", id = subject,  
                 corstr = "independence", scale.fix = TRUE, scale.value = 1)  
summary(resp_gee1)
```

GEE: GENERALIZED LINEAR MODELS FOR DEPENDENT DATA
gee S-function, version 4.13 modified 98/01/27 (1998)

Model:

Link: Logit
Variance to Mean Relation: Binomial
Correlation Structure: Independent

Call:

```
gee(formula = nstat ~ centre + treatment + sex + baseline + age,  
     id = subject, data = resp, family = "binomial", corstr = "independence",  
     scale.fix = TRUE, scale.value = 1)
```

Summary of Residuals:

Min	1Q	Median	3Q	Max
-0.93134415	-0.30623174	0.08973552	0.33018952	0.84307712

Coefficients:

	Estimate	Naive S.E.	Naive z	Robust S.E.	Robust z
(Intercept)	0.31025629	0.452642507	0.6854334	0.65618360	0.4728193
centre	0.67160098	0.239566599	2.8033999	0.35681913	1.8821889
treatment	1.29921589	0.236841017	5.4856034	0.35077797	3.7038127
sexmale	0.11924365	0.294671045	0.4046670	0.44320235	0.2690501
baselinepoor	-1.88202860	0.241290221	-7.7998545	0.35005152	-5.3764332
age	-0.01816588	0.008864403	-2.0493061	0.01300426	-1.3969169

Estimated Scale Parameter: 1

Number of Iterations: 1

Working Correlation

	[,1]	[,2]	[,3]	[,4]
[1,]	1	0	0	0
[2,]	0	1	0	0
[3,]	0	0	1	0
[4,]	0	0	0	1

GEE in R

- Naive SE should be similar to Robust SE
- In this model, they're not similar
 - Indicates incorrect correlation matrix structure
- Re-run marginal model via GEE using another type of correlation matrix structure
 - We're trying to assess ***correlations*** in the data
 - Correlation assessed by matrix covariance expressed as the off-diagonal terms
 - In this matrix, off-diagonal terms = 0

```
resp_gee1<-gee(nstat ~ centre + treatment + sex + baseline + age,  
                 data = resp, family = "binomial", id = subject,  
                 corstr = "independence", scale.fix = TRUE, scale.value = 1)  
summary(resp_gee1)
```

```
GEE: GENERALIZED LINEAR MODELS FOR DEPENDENT DATA  
gee S-function, version 4.13 modified 98/01/27 (1998)
```

Model:

Link: Logit
Variance to Mean Relation: Binomial
Correlation Structure: Independent

Call:

```
gee(formula = nstat ~ centre + treatment + sex + baseline + age,  
     id = subject, data = resp, family = "binomial", corstr = "independence",  
     scale.fix = TRUE, scale.value = 1)
```

Summary of Residuals:

Min	1Q	Median	3Q	Max
-0.93134415	-0.30623174	0.08973552	0.33018952	0.84307712

Coefficients:

	Estimate	Naive S.E.	Naive z	Robust S.E.	Robust z
(Intercept)	0.31025629	0.452642507	0.6854334	0.65618360	0.4728193
centre	0.67160098	0.239566599	2.8033999	0.35681913	1.8821889
treatment	1.29921589	0.236841017	5.4856034	0.35077797	3.7038127
sexmale	0.11924365	0.294671045	0.4046670	0.44320235	0.2690501
baselinepoor	-1.88202860	0.241290221	-7.7998545	0.35005152	-5.3764332
age	-0.01816588	0.008864403	-2.0493061	0.01300426	-1.3969169

Estimated Scale Parameter: 1

Number of Iterations: 1

Working Correlation

	[,1]	[,2]	[,3]	[,4]
[1,]	1	0	0	0
[2,]	0	1	0	0
[3,]	0	0	1	0
[4,]	0	0	0	1

GEE in R

- Z-statistic have a naïve and robust
 - Used as the standard normal random variables
 - Compare the z-statistic vs the critical value to obtain a p-value for that variable

```
resp_gee1<-gee(nstat ~ centre + treatment + sex + baseline + age,  
                 data = resp, family = "binomial", id = subject,  
                 corstr = "independence", scale.fix = TRUE, scale.value = 1)  
summary(resp_gee1)
```

GEE: GENERALIZED LINEAR MODELS FOR DEPENDENT DATA
gee S-function, version 4.13 modified 98/01/27 (1998)

Model:

Link: Logit
Variance to Mean Relation: Binomial
Correlation Structure: Independent

Call:

```
gee(formula = nstat ~ centre + treatment + sex + baseline + age,  
     id = subject, data = resp, family = "binomial", corstr = "independence",  
     scale.fix = TRUE, scale.value = 1)
```

Summary of Residuals:

Min	1Q	Median	3Q	Max
-0.93134415	-0.30623174	0.08973552	0.33018952	0.84307712

Coefficients:

	Estimate	Naive S.E.	Naive z	Robust S.E.	Robust z
(Intercept)	0.31025629	0.452642507	0.6854334	0.65618360	0.4728193
centre	0.67160098	0.239566599	2.8033999	0.35681913	1.8821889
treatment	1.29921589	0.236841017	5.4856034	0.35077797	3.7038127
sexmale	0.11924365	0.294671045	0.4046670	0.44320235	0.2690501
baselinepoor	-1.88202860	0.241290221	-7.7998545	0.35005152	-5.3764332
age	-0.01816588	0.008864403	-2.0493061	0.01300426	-1.3969169

Estimated Scale Parameter: 1

Number of Iterations: 1

Working Correlation

	[,1]	[,2]	[,3]	[,4]
[1,]	1	0	0	0
[2,]	0	1	0	0
[3,]	0	0	1	0
[4,]	0	0	0	1

GEE in R

- One more thing before we try another correlation matrix structure...
- Can round each estimate to a specific number of significant digits to make interpretation a bit easier

```
> round(summary(resp_gee1)$coefficients,4)
            Estimate Naive S.E. Naive z Robust S.E. Robust z
(Intercept) 0.3103   0.4526  0.6854   0.6562  0.4728
centre       0.6716   0.2396  2.8034   0.3568  1.8822
treatment    1.2992   0.2368  5.4856   0.3508  3.7038
sexmale      0.1192   0.2947  0.4047   0.4432  0.2691
baseline     -1.8820   0.2413 -7.7999   0.3501 -5.3764
age          -0.0182   0.0089 -2.0493   0.0130 -1.3969
> round(summary(resp_gee1)$working.correlation,3)
      [,1] [,2] [,3] [,4]
[1,]    1    0    0    0
[2,]    0    1    0    0
[3,]    0    0    1    0
[4,]    0    0    0    1
```

GEE in R

- Designated correlation matrix structure (“corstr”) as *exchangeable*
- Naïve SE now much closer to Robust SE
 - Suggests correlation matrix structure (exchangeable) is the optimal structure for this dataset
 - Correlation matrix showing off-diagonal terms (covariance) estimates of 0.3359

```
resp_gee2<-gee(nstat ~ centre + treatment + sex + baseline + age,  
                 data = resp, family = "binomial", id = subject,  
                 corstr = "exchangeable", scale.fix = TRUE, scale.value = 1)  
summary(resp_gee2)
```

GEE: GENERALIZED LINEAR MODELS FOR DEPENDENT DATA
gee S-function, version 4.13 modified 98/01/27 (1998)

Model:

Link:

Logit

Variance to Mean Relation: Binomial

Correlation Structure: Exchangeable

Call:

```
gee(formula = nstat ~ centre + treatment + sex + baseline + age,  
     id = subject, data = resp, family = "binomial", corstr = "exchangeable",  
     scale.fix = TRUE, scale.value = 1)
```

Summary of Residuals:

Min	1Q	Median	3Q	Max
-0.93134415	-0.30623174	0.08973552	0.33018952	0.84307712

Coefficients:

	Estimate	Naive S.E.	Naive z	Robust S.E.	Robust z
(Intercept)	0.31025629	0.6414066	0.4837124	0.65618360	0.4728193
centre	0.67160098	0.3394723	1.9783676	0.35681913	1.8821889
treatment	1.29921589	0.3356101	3.8712064	0.35077797	3.7038127
sexmale	0.11924365	0.4175568	0.2855747	0.44320235	0.2690501
baselinepoor	-1.88202860	0.3419147	-5.5043802	0.35005152	-5.3764332
age	-0.01816588	0.0125611	-1.4462014	0.01300426	-1.3969169

Estimated Scale Parameter: 1

Number of Iterations: 1

Working Correlation

	[,1]	[,2]	[,3]	[,4]
[1,]	1.0000000	0.3359883	0.3359883	0.3359883
[2,]	0.3359883	1.0000000	0.3359883	0.3359883
[3,]	0.3359883	0.3359883	1.0000000	0.3359883
[4,]	0.3359883	0.3359883	0.3359883	1.0000000

GEE in R

- Rounded figures....

	Estimate	Naive S.E.	Naive z	Robust S.E.	Robust z
(Intercept)	0.3103	0.6414	0.4837	0.6562	0.4728
centre	0.6716	0.3395	1.9784	0.3568	1.8822
treatment	1.2992	0.3356	3.8712	0.3508	3.7038
sexmale	0.1192	0.4176	0.2856	0.4432	0.2691
baselinepoor	-1.8820	0.3419	-5.5044	0.3501	-5.3764
age	-0.0182	0.0126	-1.4462	0.0130	-1.3969


```
> round(summary(resp_gee2)$working.correlation,3)
      [,1]  [,2]  [,3]  [,4]
[1,] 1.000 0.336 0.336 0.336
[2,] 0.336 1.000 0.336 0.336
[3,] 0.336 0.336 1.000 0.336
[4,] 0.336 0.336 0.336 1.000
```

Hypothesis Testing

- Calculate p-values to determine which covariates are significant predictors of counts of binary response data
 - Reminder: parameter estimates are the *log odds ratio* for that variable/covariate

```
##### Hypothesis testing-----
teststat.gee2<-round(summary(resp_gee2)$coefficients,4)
#### calculate p-value -----
pval<-round(pnorm(abs(teststat.gee2[,5])),lower.tail=FALSE)^2,4)
## final table-----
cbind(teststat.gee2,pval)
```

	Estimate	Naive S.E.	Naive z	Robust S.E.	Robust z	pval
(Intercept)	0.3103	0.6414	0.4837	0.6562	0.4728	0.6364
centre	0.6716	0.3395	1.9784	0.3568	1.8822	0.0598
treatment	1.2992	0.3356	3.8712	0.3508	3.7038	0.0002
sexmale	0.1192	0.4176	0.2856	0.4432	0.2691	0.7879
baselinepoor	-1.8820	0.3419	-5.5044	0.3501	-5.3764	0.0000
age	-0.0182	0.0126	-1.4462	0.0130	-1.3969	0.1624

Final Marginal Model

$$\text{Logit}(P(Y_{ij} = \text{good})) = \beta_0 + \beta_1(\text{center}) + \beta_2(\text{Treatment | DrugA}) + \beta_3(\text{Baseline | Poor})$$

```
resp_gee3<-gee(nstat ~ centre + treatment + baseline,
  data = resp, family = "binomial", id = subject,
  corstr = "exchangeable", scale.fix = TRUE, scale.value = 1)
summary(resp_gee3)
```

	Estimate	Naive S.E.	Naive z	Robust S.E.	Robust z
(Intercept)	-0.0803	0.5828	-0.1378	0.5743	-0.1398
centre	0.5616	0.3217	1.7455	0.3363	1.6699
treatment	1.2912	0.3289	3.9259	0.3274	3.9435
baselinepoor	-1.9105	0.3423	-5.5811	0.3412	-5.5985

```
> round(summary(resp_gee3)$working.correlation,3)
 [,1] [,2] [,3] [,4]
[1,] 1.000 0.349 0.349 0.349
[2,] 0.349 1.000 0.349 0.349
[3,] 0.349 0.349 1.000 0.349
[4,] 0.349 0.349 0.349 1.000
```

Does Drug have a Significant Effect on Response?

- Exponentiate the “treatment” coefficient (β ; parameter estimate) to get the OR
 - Calculate 95% CI around that *estimate*

```
### interpretation of population averages through odds ratios---
##### Oddsratio and confidence interval-----
se<-summary(resp_gee3)$coefficients["treatmenttreatment","Robust S.E."]
ci<-exp(coef(resp_gee3)["treatmenttreatment"] + c(-1,1) * se * qnorm(0.975))
cbind(OR=exp(coef(resp_gee3)["treatmenttreatment"]),ci)
```

	OR	ci
[1,]	3.637184	1.914516
[2,]	3.637184	6.909896

INTERPRETATION:

- The **odds** of an **average person** achieving a “good” respiratory status with DrugA is 3.64x more than an average person taking placebo
 - A typical person taking DrugA has between **2x – 7x** better odds of achieving good respiratory status than a typical person taking placebo

Does Baseline status have a Significant Effect on Response?

```
##### Oddsratio and confidence interval-----
sel<-summary(resp_gee3)$coefficients["baselinepoor","Robust S.E."]
ci1<-exp(coef(resp_gee3)["baselinepoor"] + c(-1,1) * sel * qnorm(0.975))
cbind(OR=exp(coef(resp_gee3)["baselinepoor"]),ci1)
```

	OR	ci1
	0.1480121	0.07582806
	0.1480121	0.28891109

INTERPRETATION:

- The **odds** of an ***average person*** achieving a “good” respiratory status with a poor baseline respiratory status is only 0.15x more than an average person with a good baseline status (ranging between **0.08 – 0.30x**)

Model Performance

- Generalized linear models (GLM), like those used for logistic and poisson regression, using maximum likelihood estimation (MLE) algorithms to estimate parameters
 - Provides each model with an AIC value
 - Can compare models run with MLE by AIC
 - The lower the AIC, the better the model (among those models using MLE)
- Marginalized Models use GEE to estimate parameters (not MLE)
 - Doesn't provide an AIC to evaluate model performance and determine "final" models
 - How do we compare model performance among those run with GEE?

Model Performance

- GEE is not a likelihood-based method
- However, can use a Quasi-likelihood under Independence model Criterion (QIC)
 - QIC can be used to find acceptable working correlation structure for a given model
- The QIC_u statistic can be used to compare models run with GEE
 - i.e. QIC_u can be used for model selection

2. Generalized Linear Mixed Effects Modeling

- Subject-specific model
 - Inference made at the individual subject level
- Model the mean response
 - As a function of covariates
 - Include random effects (intra- and inter-subject differences)
 - Random effects follow a specific distribution
 - Inclusion of the random effects accounts for correlation between observations

GLMM

- Because a “subject-specific” model, GLMM estimates between-subject variability (BSV) through the use of random effects
- Conditional on the random effects, the responses from each subject are independent and are assumed to follow the exponential family distribution
- Since a “generalized linear *mixed-effects* model”, has same 3 components as generalized linear models
 1. Distributional assumption
 2. Systematic component
 3. Link function

GLMM Distributional Assumption

$Y_{ij} | \eta_{ij} \sim \text{Exponential family distribution}$

- response for a specific subject (j) at a specific time point (i) is a function of the between subject variability (BSV) that is estimated via random effects parameter (η_{ij})
- If the response is longitudinal **binary**, then:

$Y_{ij} | \eta_{ij} \sim \text{Bernoulli distribution, where: } \text{Var}(Y_{ij} | \eta_{ij}) = E(Y_{ij} | \eta_{ij})(1 - E(Y_{ij} | \eta_{ij}))$

- If the response is longitudinal **count** data, then:

$\text{Var}(Y_{ij} | \eta_{ij}) = E(Y_{ij} | \eta_{ij})$ (Poisson distribution)

Systematic Component and Link Function

$$g(E(Y_{ij} | \eta_i)) = X_{ij}\beta + Z_{ij}\eta_i$$

Population mean

Fixed effect

Subject-specific

Random effect

What differentiates that specific person from a “typical” person (aka population mean)

Subject-specific mean

or

Conditional mean response

What the mean parameter estimate is for that subject

$$\log\left(\frac{P(Y_{ij} = 1) | \eta_i}{1 - P(Y_{ij} = 1 | \eta_i)}\right) = X_{ij}\beta + Z_{ij}\eta_i$$

GLMM model for longitudinal binary responses

Fixed effects + random effects = mixed effects

GLMM: Generalized linear **mixed effects** model

Random Effects

- Random effects (subject-specific differences from population mean) are assumed to follow a multi-variate normal distribution
 - $\eta_i \sim MVN(0, \Omega)$
 - read as the random effect for subject i is proportional to the multivariate normal distribution, with a center of 0 and variance
 - a **negative** random effect for subject i means the mean for subject i for that parameter is **below** the population mean
 - a **positive** random effect for subject i means the mean for subject i for that parameter is **above** the population mean

Random Effects

- Intercept-only model ($Z_{ij} = 1$):

$$\log\left(\frac{P(Y_{ij} = 1) | \eta_i}{1 - P(Y_{ij} = 1 | \eta_i)}\right) = X_{ij}\beta + \eta_i$$

- $\eta_i \sim N(0, \omega_{11})$, where ω_{11} is a univariate variance term

- Slope + Intercept model:

$$\log\left(\frac{P(Y_{ij}=1) | \eta_i}{1-P(Y_{ij}=1 | \eta_i)}\right) = X_{ij}\beta + Z_{ij}\eta_i , \text{ where } Z_{ij} = (1, t_{ij}), \text{ and } t_{ij} \text{ is a predictor variable wrt time}$$

$$\eta_i \sim N(0, \begin{bmatrix} \omega_{11} & \omega_{12} \\ \omega_{21} & \omega_{22} \end{bmatrix})$$

- two parameters for random effects (Z_{ij}, t_{ij}), thus a bivariate variance term
 - need to construct a 2×2 variance-covariance matrix
 - diagonals are variance, off-diagonals are covariance between the two parameters

Estimation and Inference

- GLMM uses MLE (an iterative process) to estimate optimal parameter estimates
- Uses a joint distribution of vector responses and a vector of random effects, both of which are fully specified
- Comparing between models can be done by likelihood ratio test (LRT) or comparing AIC from each model

GLMM using R

- Base model (intercept-only)
- Random effects (“(1|subject)”)
- Model diagnostics include AIC, deviance, and log likelihood
- Only fixed effect is the intercept
- Subject-specific random effects only around intercept

```
#### Generalized linear mixed effects model (GLMM)---  
install.packages("lme4")  
library(lme4)  
resp.glmm0<-glmer(nstat ~ 1 + (1|subject),  
                     data = resp, family = binomial, REML=F )  
summary(resp.glmm0)
```

```
Generalized linear mixed model fit by maximum likelihood (Laplace ,  
glmerMod)  
Family: binomial ( logit )  
Formula: nstat ~ 1 + (1 | subject)  
Data: resp
```

AIC	BIC	logLik	deviance	df.resid
492.7	500.9	-244.3	488.7	442

Scaled residuals:

Min	1Q	Median	3Q	Max
-1.6536	-0.3281	0.2797	0.2797	1.5336

Random effects:

Groups	Name	Variance	Std.Dev.
subject	(Intercept)	7.035	2.652

Number of obs: 444, groups: subject, 111

Fixed effects:

	Estimate	Std. Error	z value	Pr(> z)
(Intercept)	0.5053	0.2912	1.735	0.0827 .

Signif. codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

“REML = F” tells glmer function to use Maximum Likelihood Estimation (MLE) instead of Restricted MLE (aka REML)

GLMM using R

- Full model
- AIC, loglik, and deviance all lower
- Treatment (DrugA) and poor baseline status are both significant predictors
 - Center is borderline significant

```
resp.glmm<-glmer(nstat ~ centre + treatment + baseline+sex+age+(1|subject),  
                   data = resp, family = binomial, REML=F )  
summary(resp.glmm)
```

```
> summary(resp.glmm)  
Generalized linear mixed model fit by maximum likelihood (Laplace Approximation) [  
glmerMod]  
  Family: binomial ( logit )  
Formula: nstat ~ centre + treatment + baseline + sex + age + (1 | subject)  
  Data: resp
```

AIC	BIC	logLik	deviance	df.resid
443.0	471.7	-214.5	429.0	437

Scaled residuals:

Min	1Q	Median	3Q	Max
-2.7832	-0.3652	0.1428	0.3735	2.1801

Random effects:

Groups	Name	Variance	Std.Dev.
subject	(Intercept)	3.86	1.965
Number of obs:	444, groups:	subject,	111

Fixed effects:

	Estimate	Std. Error	z value	Pr(> z)
(Intercept)	0.38176	1.03804	0.368	0.713045
centre	1.04332	0.54605	1.911	0.056047
treatment	2.15611	0.55444	3.889	0.000101 ***
baselinepoor	-3.06824	0.60032	-5.111	3.21e-07 ***
sexmale	0.20180	0.67054	0.301	0.763450
age	-0.02539	0.02008	-1.265	0.205950

Signif. codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Correlation of Fixed Effects:

	(Intr)	centre	trtmnt	bslnpr	sexmal
centre	-0.680				
trtmnt	-0.184	0.058			
baselinepor	-0.364	0.151	-0.300		
sexmale	0.129	-0.147	0.219	-0.102	
age	-0.411	-0.223	-0.050	-0.016	-0.263

$$\text{Logit}(P(Y_{ij} = \text{good})) = \beta_0 + \beta_1(\text{trt}=\text{DrugA}) + \beta_2(\text{baseline}=\text{poor}) + \eta_i$$

Final Model

- Final model
- AIC, loglik, and deviance all lower
- Treatment (DrugA) and poor baseline status are both significant predictors
 - Center is borderline significant

```
### removed age, gender, center from model to get final model--
resp.glmm1<-glmer(nstat ~ treatment + baseline+(1|subject),
                     data = resp, family = binomial, REML=F )
summary(resp.glmm1)
```

Family: binomial (logit)
 Formula: nstat ~ treatment + baseline + (1 | subject)
 Data: resp

AIC	BIC	logLik	deviance	df.resid
441.6	458.0	-216.8	433.6	440

Scaled residuals:

Min	1Q	Median	3Q	Max
-2.4275	-0.3756	0.1453	0.3631	1.9619

Random effects:

Groups	Name	Variance	Std.Dev.
subject	(Intercept)	4.149	2.037
Number of obs: 444, groups: subject, 111			

Fixed effects:

	Estimate	Std. Error	z value	Pr(> z)
(Intercept)	1.3477	0.4540	2.968	0.00299 **
treatment	2.1674	0.5527	3.921	8.80e-05 ***
baselinepoor	-3.4221	0.6125	-5.587	2.31e-08 ***

Signif. codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Correlation of Fixed Effects:

(Intr) trtmnt

trtmnt -0.312

baselinepor -0.633 -0.310

Odds Ratios



```
exp(fixef(resp.glmm1))
```

(Intercept)	treatment
3.84862425	8.73549416

baselinepoor
0.03264248

Comparing GEE vs GLMM

GEE

- Final marginal model had center, baseline status, and treatment status as significant predictors
- Exponentiated coefficients (odds ratios) represented odds of a “typical person”

GLMM

- Final subject-specific model had baseline status and treatment status as significant predictors
 - Not center
- Exponentiated coefficients (fixed effect) is the “population average” odds ratio
- Random effects variance incorporated too

To properly compare to GEE, must re-run GLMM to add center
As a covariate (even though not $p < 0.05$)

Comparing GEE vs GLMM

```
### removed age, gender from model to get final model--  
resp.glmm2<-glmer(nstat ~ centre + treatment + baseline+(1|subject),  
                     data = resp, family = binomial, REML=F )  
summary(resp.glmm2)
```

Scaled residuals:

Min	1Q	Median	3Q	Max
-2.5877	-0.3582	0.1296	0.3864	2.0640

Random effects:

Groups	Name	Variance	Std.Dev.
subject	(Intercept)	4	2

Number of obs: 444, groups: subject, 111

Fixed effects:

	Estimate	Std. Error	z value	Pr(> z)
(Intercept)	-0.1640	0.9573	-0.171	0.8639
centre	0.9068	0.5260	1.724	0.0847 .
treatment	2.1730	0.5483	3.963	7.40e-05 ***
baselinepoor	-3.1501	0.6051	-5.206	1.93e-07 ***

Signif. codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Correlation of Fixed Effects:

	(Intr)	centre	trtmnt
centre	-0.885		
trtmnt	-0.233	0.099	
baselinepor	-0.406	0.130	-0.290

> |

#####Compare gee, glmm and glm fits-----

```
out.compare<-round(cbind(exp(coef(resp_gee3)), exp(fixef(resp.glmm2))),3)  
rownames(out.compare)<-names(coef(resp_gee3))  
colnames(out.compare)<-c('GEE','GLMM')  
out.compare
```

> out.compare

	GEE	GLMM
(Intercept)	0.923	0.849
centre	1.753	2.476
treatment	3.637	8.785
baselinepoor	0.148	0.043

Comparing GEE vs GLMM

- Coefficients (β) for each predictor variable (x) are much different in GLMM vs GEE
- Why? Because coefficients have different interpretations in GLMM vs GEE
 - Hence, they're NOT comparable

```
#####Compare gee, glmm and glm fits-----
out.compare<-round(cbind(exp(coef(resp_gee3)), exp(fixef(resp.glmm2))),3)
rownames(out.compare)<-names(coef(resp_gee3))
colnames(out.compare)<-c('GEE','GLMM')
out.compare
```

```
> out.compare
```

	GEE	GLMM
(Intercept)	0.923	0.849
centre	1.753	2.476
treatment	3.637	8.785
baselinepoor	0.148	0.043

Comparing GEE vs GLMM

	GEE	GLMM
(Intercept)	0.923	0.849
centre	1.753	2.476
treatment	3.637	8.785
baselinepoor	0.148	0.043
.	.	.

GEE

Treatment effect:

- The odds of an ***average person*** achieving a “good” respiratory status when taking Drug A is 3.64x better than the odds of an ***average person*** taking placebo

GLMM

Treatment effect:

- The odds of a ***specific person*** achieving a “good” respiratory status when taking Drug A is 8.79x better than the odds of that **same specific person** taking placebo

PM Break

Exposure/Response Modeling – Dose Determination

- A pharmaceutical company is developing a new analgesic to be used for post-operative pain. A dose-ranging (phase I) study has been conducted in 160 patients, and you have been asked to provide input on the dose to be carried forward to phase II trials and beyond
- Drug tested at 3 dose levels (5 mg QD, 20 mg QD, and 80 mg QD)
- Placebo controlled trial
- Have drug conc vs time data (0-8hr post dose)
- Longitudinal binary response (pain relief: no=0; yes=1)
 - Response at each time point that PKs were taken

Primary Objective

- Select the optimal dose of drug for future clinical trials to give the drug the best chance to demonstrate significant efficacy with manageable off-target (side) effects

Assumptions

1. That all 160 patients had 100% compliance
 - i.e. they each took the drug when they were told, and that PK time points post-dose are accurate of the actual drug disposition
2. All patients had normal organ (specifically liver and kidney) function
3. All patients had the same scale of “pain threshold”
 - Patients could have differing levels of what they felt was “pain” to report as pain yes/no
 - Assume that each patient abided by same scale

Methods

- Examine dataset via GEE for dose selection
 - Available covariates: drug concentration (at different times), treatment arm (placebo, 5mg Drug, 20mg Drug, 80 mg Drug)
 - Provide interpretations of population-averaged odds ratios for each potential predictor variable
- Examine dataset via GLMM for dose selection
 - Factor in subject-specific variability (random effects) onto the “population estimate” (fixed effects) for each predictor variable
 - Provide subject-specific odds of experiencing pain relief

Dataset Exploration

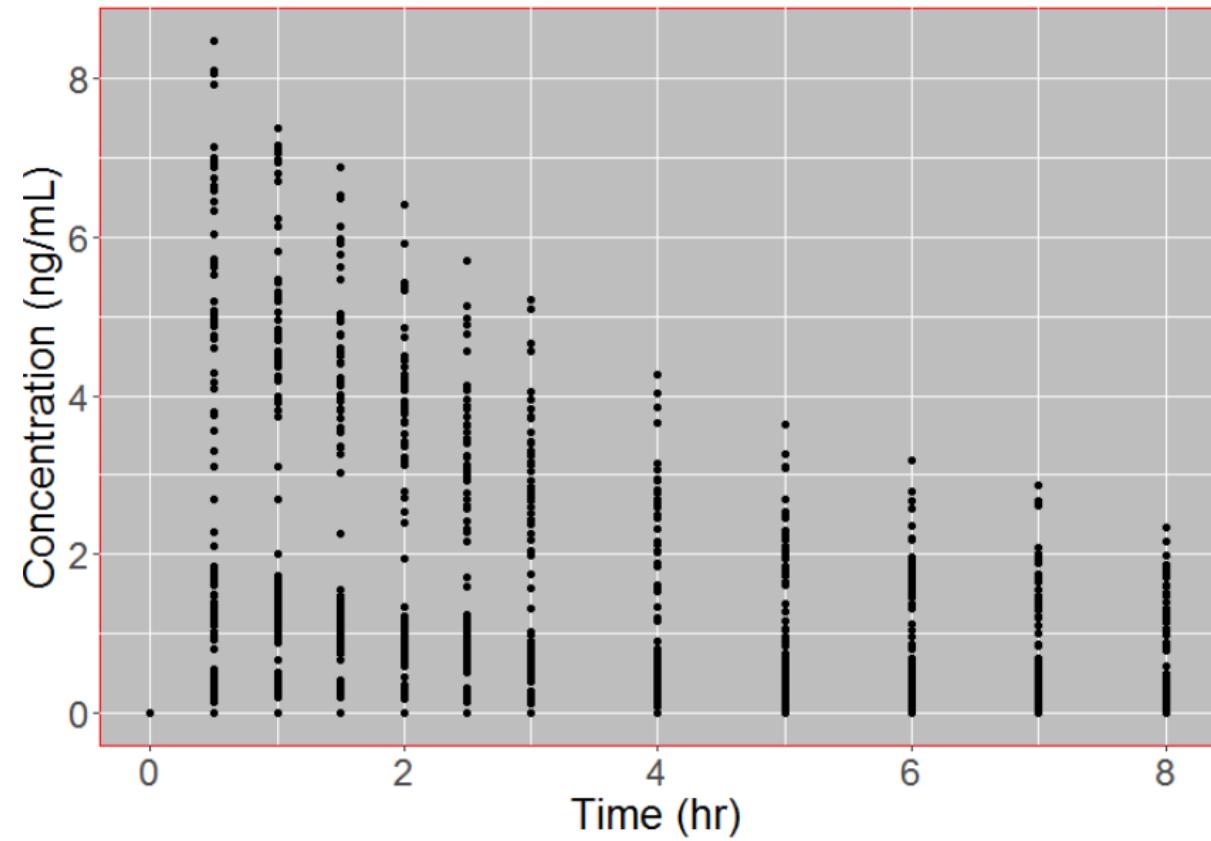
```
pain<-read.csv("Painrelief.csv", sep=",")  
summary(pain)
```

```
> summary(pain)  
      ARM          ID          TIME         CONC        PAINRELIEF       DOSE  
A20_0_at2h:480  Min.   : 1.00  Min.   :0.000  Min.   :0.0000  Min.   :0.0000  Min.   : 0.00  
A5_0_at2h :480  1st Qu.:40.75  1st Qu.:1.375  1st Qu.:0.0000  1st Qu.:0.0000  1st Qu.: 3.75  
A80_0_at2h:480 Median :80.50  Median :2.750  Median :0.2287  Median :0.0000  Median :12.50  
Placebo  :480  Mean   :80.50  Mean   :3.375  Mean   :0.9302  Mean   :0.4974  Mean   :26.25  
                  3rd Qu.:120.25 3rd Qu.:5.250  3rd Qu.:1.1514  3rd Qu.:1.0000  3rd Qu.:35.00  
                  Max.   :160.00  Max.   :8.000  Max.   :8.4719  Max.   :1.0000  Max.   :80.00
```

Exploratory Plots – Exposure (PK)

- Appears that maximum drug concentration reached at 30min (0.5 hr) post oral dose
- NOT subsetted by dose

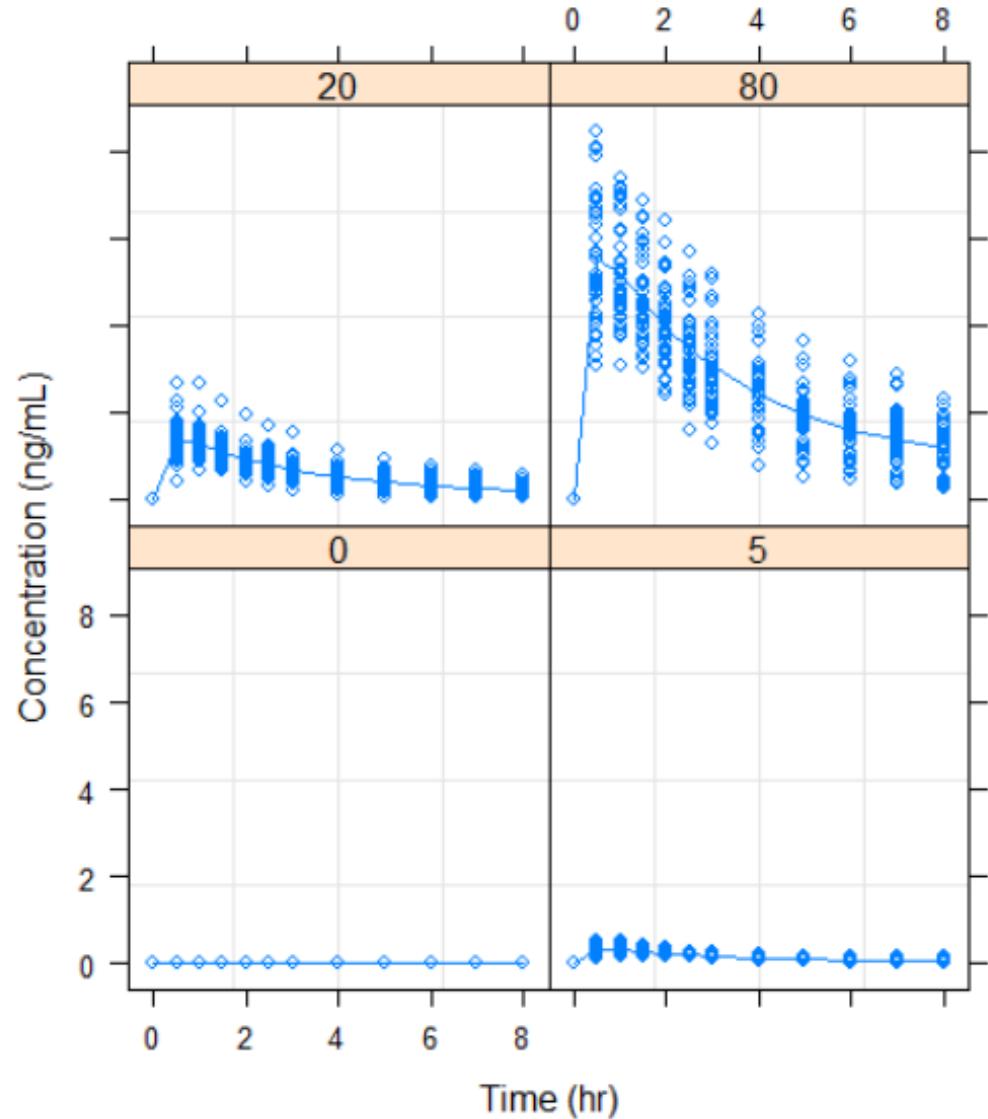
```
##### Conc vs time curve, all together---  
library(ggplot2)  
plot1<-ggplot(data = pain, aes(x = TIME, y = CONC, group=ARM))+geom_point() +  
  xlab("Time (hr)") +  
  ylab("Concentration (ng/mL)")  
plot1+theme(panel.background = element_rect(fill='grey', colour='red')) +  
  theme(axis.title.y = element_text(colour = 'black', size = 20)) +  
  theme(axis.text.y = element_text(size = 18)) +  
  theme(axis.title.x = element_text(colour = 'black', size = 20)) +  
  theme(axis.text.x = element_text(size = 18)) +  
  theme(plot.title = element_text(lineheight=.8, face="bold")) +  
  theme(legend.position="top") +  
  theme(legend.title=element_blank()) +  
  theme(legend.text = element_text(colour="blue", size = 18, face = "bold"))
```



Exploratory Plots – Exposure (PK)

- Appears that maximum drug concentration reached at 30min (0.5 hr) post oral dose
- Subsetted by dose
- Concentrations appear to increase with dose amount

```
#### group data by DOSE and plot ####
install.packages("nlme")
library(nlme)
pain.2<-groupedData(CONC~TIME | DOSE, data=pain,
                      labels = list(x="Time(hr)", y="Conc(ng/mL)"))
plot2<-plot(pain.2, aspect=1/1, ylab="Concentration (ng/mL)",
            xlab="Time (hr)")
plot2
```



Exploratory Plots – Response (PD)

Q. Does the number (count) of positive pain relief (Y=1) increase with dose?

```
> table(pain$PAINRELIEF, pain$DOSE)
```

	0	5	20	80
0	344	266	194	161
1	136	214	286	319

A. Yes

Exploratory Plots – Response (PD)

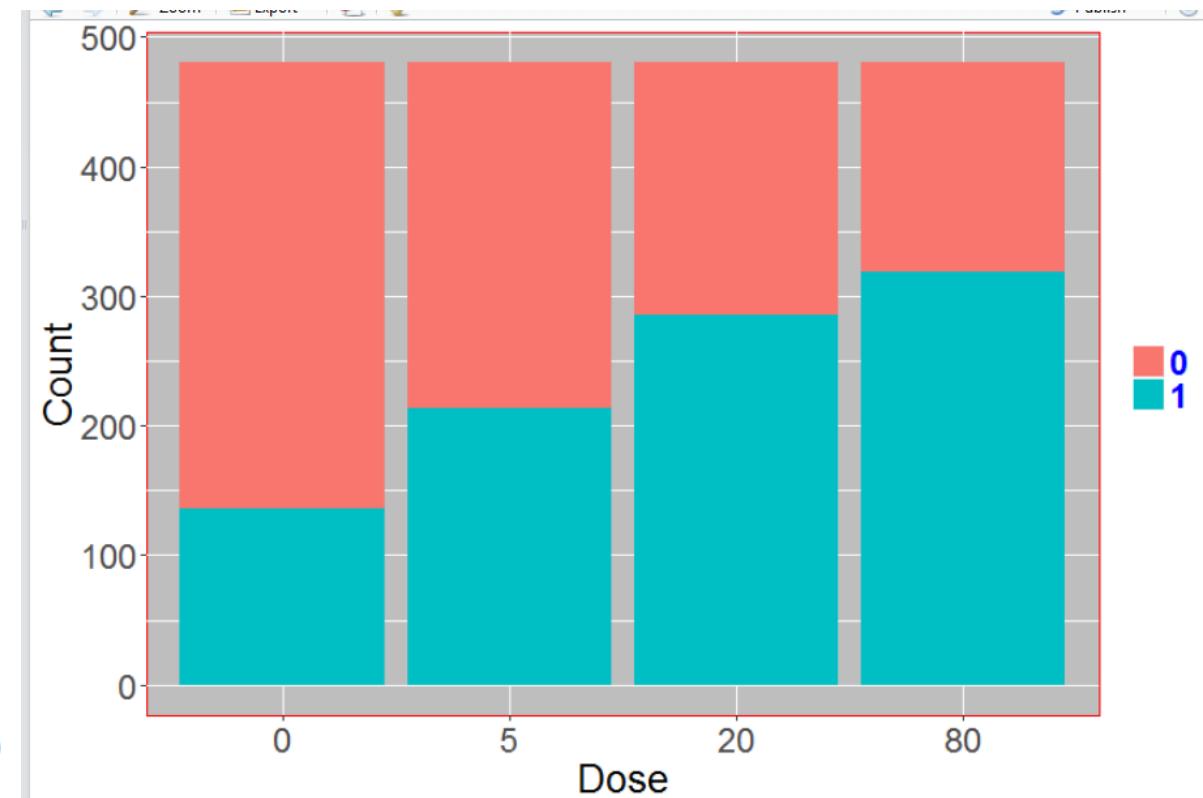
Depict in a figure....

```
##### Re-arrange data for stacked bar chart
library(reshape2)
count1<-ftable(xtabs(~PAINRELIEF+DOSE,data=pain))
count1

### melt fnc converts data from a wide format to a
mydf = melt(count1)
names(mydf) = c("PainRelief", "Dose", "Count")
head(mydf)

###plot it--
##### Stacked bar chart-----
library(reshape2)
plot4<-ggplot(mydf, aes(x= Dose, y= Count, fill=PainRelief)) +
  geom_bar(stat="identity")

plot4+ theme(panel.background = element_rect(fill='gray', colour='red'))+
  theme(axis.title.y = element_text(colour = 'black', size = 20))+  
theme(axis.text.y = element_text(size = 18))+  
theme(axis.title.x = element_text(colour = 'black', size = 20))+  
theme(axis.text.x = element_text(size = 18))+  
theme(plot.title = element_text(lineheight=.8, face="bold"))+  
theme(legend.position="right")+
  theme(legend.title=element_blank())+
  theme(legend.text = element_text(colour="blue", size = 18, face = "bold"))+
  theme(strip.text.x = element_text(size = 12, face = "bold",colour = "black"))
```



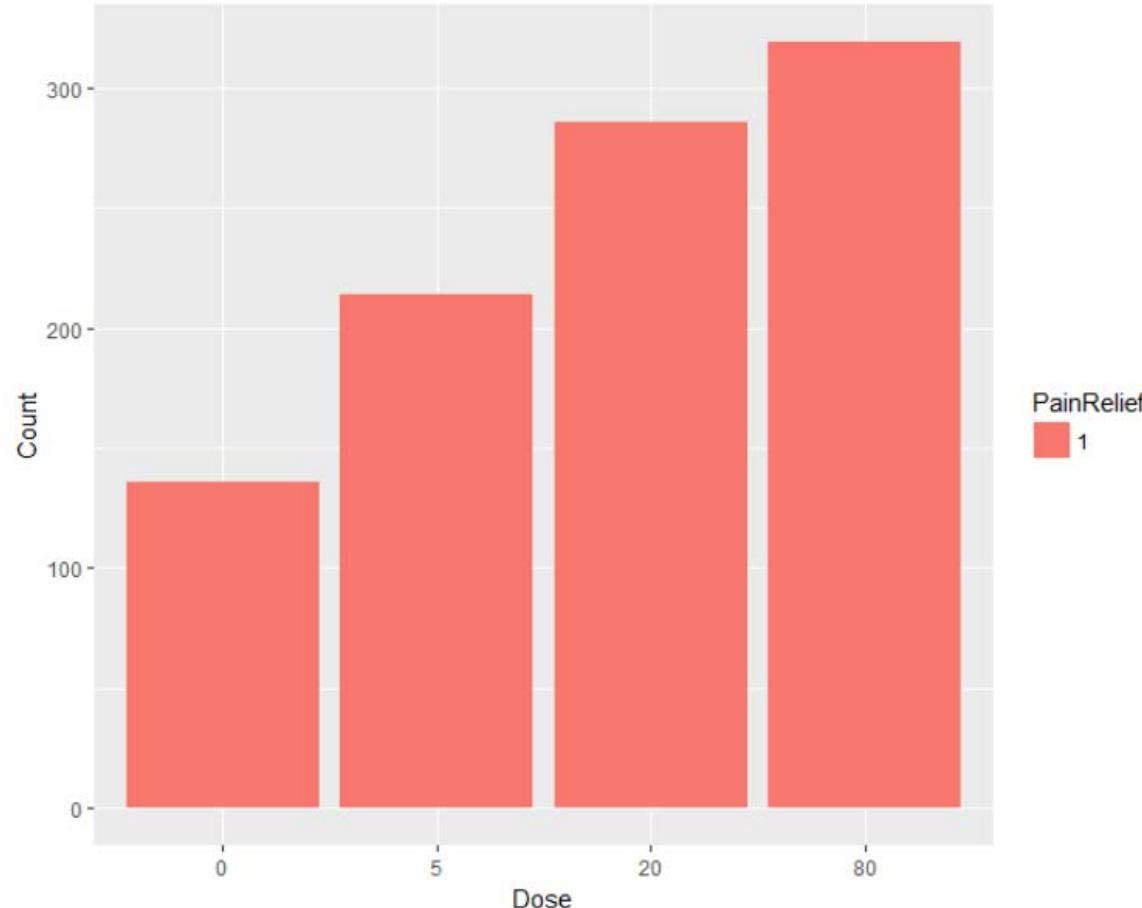
Exploratory Plots – Response (PD)

Only looking at successfully pain relief responses

There seems to be a plateau effect in response with dose

```
##### Using only successfully responses #####
### only PainRelief=1 responses###
nopain<-pain[pain$PAINRELIEF==1,]
count3<-ftable(xtabs(~PAINRELIEF+DOSE,data=nopain))
### melt fnc converts data from a wide format to a
mydf3 = melt(count3)
names(mydf3) = c("PainRelief", "Dose", "Count")
head(mydf3)

##### Stacked bar chart-----
ggplot(mydf3, aes(x= Dose, y= Count, fill=PainRelief)) +
  geom_bar(stat="identity")
```



Preliminary Observations

- There appears to be a dose-dependent increase in **exposure**
- There appears to be a dose-dependent increase in **response**
- There likely IS a significant drug effect (exposure/response)
- What is the tentative time course of drug effect?
 - We know Cmax occurs at 30 min post dose
 - When dose maximum pain relief YES response (Rmax) occur?

Exploratory Plots – Response over Time

Q. When does Rmax occur relative to Cmax?

```
> ftable(xtabs(~DOSE+PAINRELIEF+TIME,data=pain))
```

		DOSE	PAINRELIEF	TIME									
0	0.5			0	0.5	1	1.5	2	2.5	3	4	5	6
0	0			38	30	25	27	25	26	25	29	29	31
	1			2	10	15	13	15	14	15	11	11	9
5	0			36	24	17	17	14	21	19	22	23	19
	1			4	16	23	23	26	19	21	18	17	21
20	0			37	16	16	11	13	13	9	11	20	16
	1			3	24	24	29	27	27	31	29	20	24
80	0			38	12	9	9	8	8	9	12	12	11
	1			2	28	31	31	32	32	31	28	28	29

A. Appears to be dose-dependent, but somewhere between 1.5 – 2 hr post dose

Exploratory Plots – Response over Time

Depict in a figure....

Plot confirms Rmax occurs at 2 hr post dose
- a 1.5 hr delay in response from exposure max

```
##### count of pain relief by time #####
count2<-ftable(xtabs(~PAINRELIEF+TIME,data=pain))
count2
mydf2 = melt(count2)
names(mydf2) = c("PainRelief", "Time", "Count")
mydf2

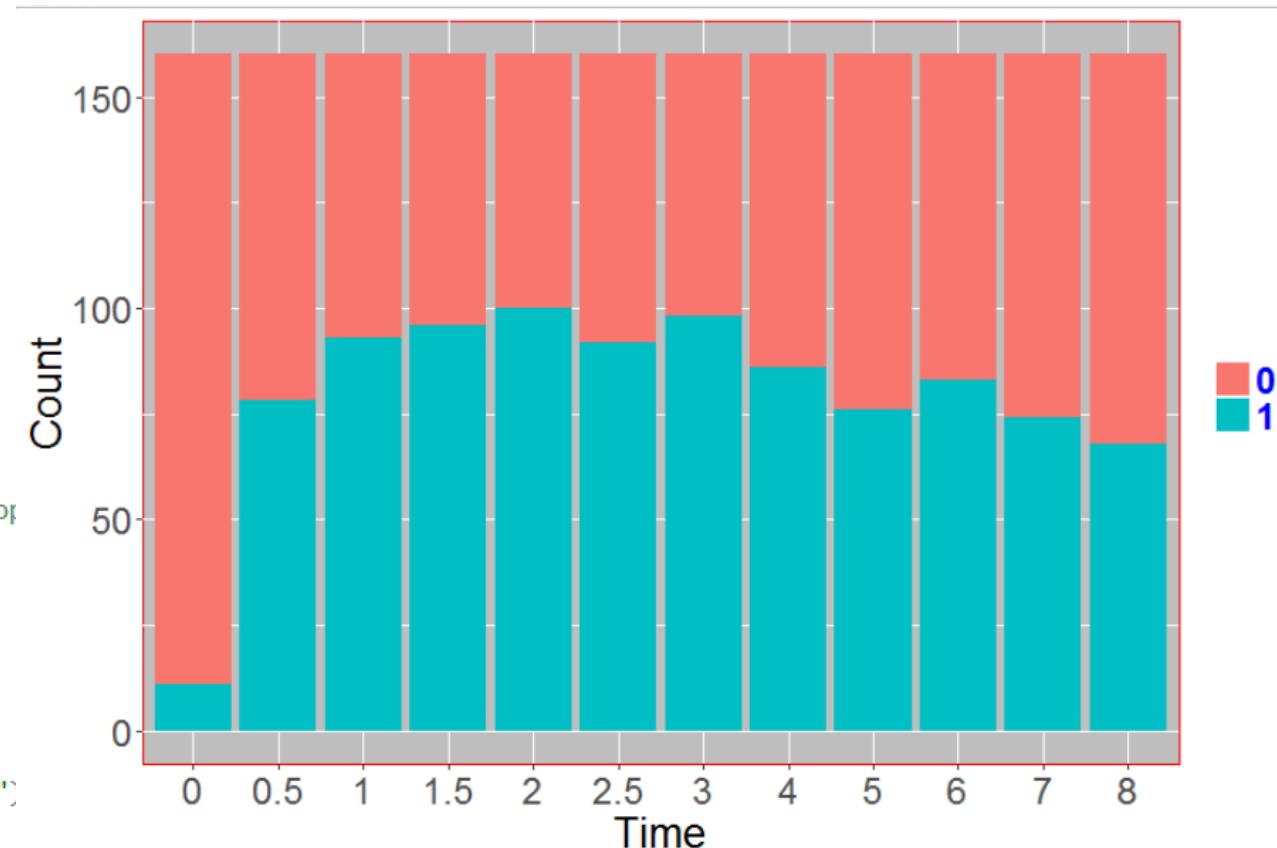
plot5<-ggplot(mydf2, aes(x= Time, y= Count, fill=PainRelief)) +
  geom_bar(stat="identity")
##### plot attributes - axis labels, axis titles, background, ##### legend #####
plot5+ theme(panel.background = element_rect(fill='gray', colour='red'))+
  theme(axis.title.y = element_text(colour = 'black', size = 20))+  

  theme(axis.text.y = element_text(size = 18))+  

  theme(axis.title.x = element_text(colour = 'black', size = 20))+  

  theme(axis.text.x = element_text(size = 18))+  

  theme(plot.title = element_text(lineheight=.8, face="bold"))+
  theme(legend.position="right")+
  theme(legend.title=element_blank())+
  theme(legend.text = element_text(colour="blue", size = 18, face = "bold"))
```



Exploratory Plots – Response over Time

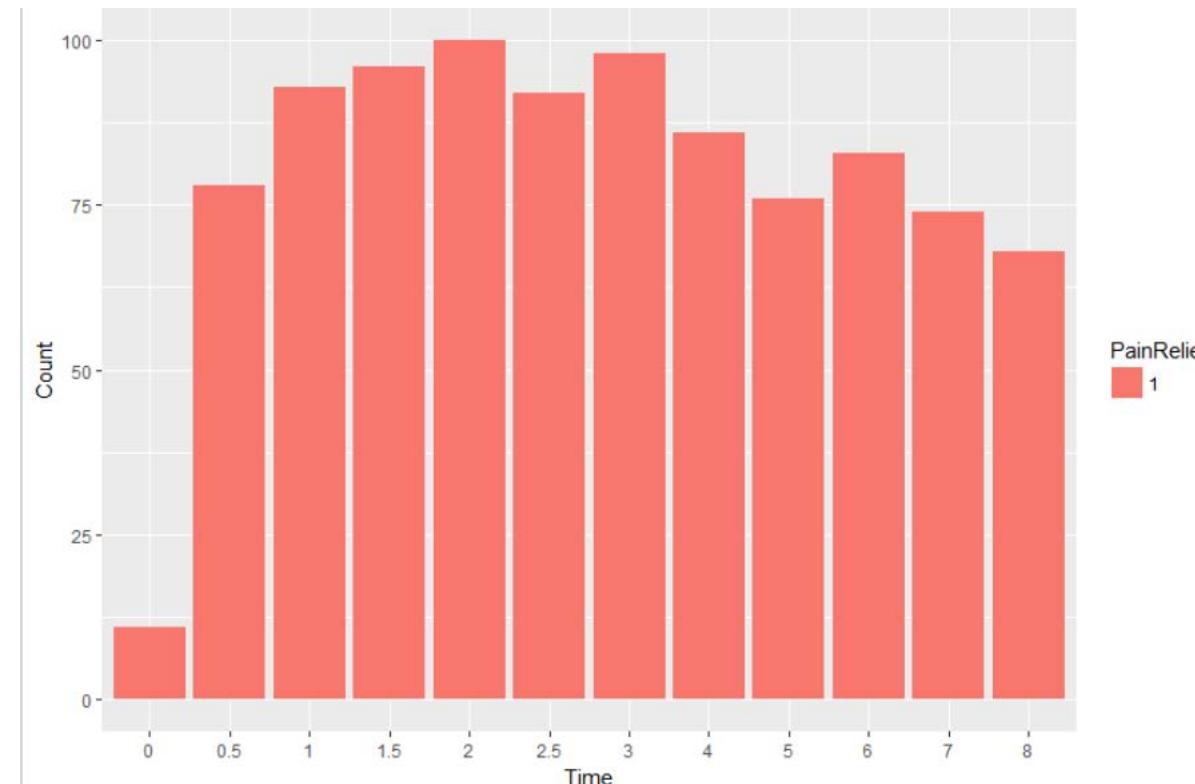
Look at only successful pain relief (response = 1; yes)

Subset out only painrelief=1

```
##### Using only successfully responses #####
### only PainRelief=1 responses###
nopain<-pain[pain$PAINRELIEF==1,]
count3<-ftable(xtabs(~PAINRELIEF+DOSE,data=nopain))
### melt fnc converts data from a wide format to a
mydf3 = melt(count3)
names(mydf3) = c("PainRelief", "Dose", "Count")
head(mydf3)

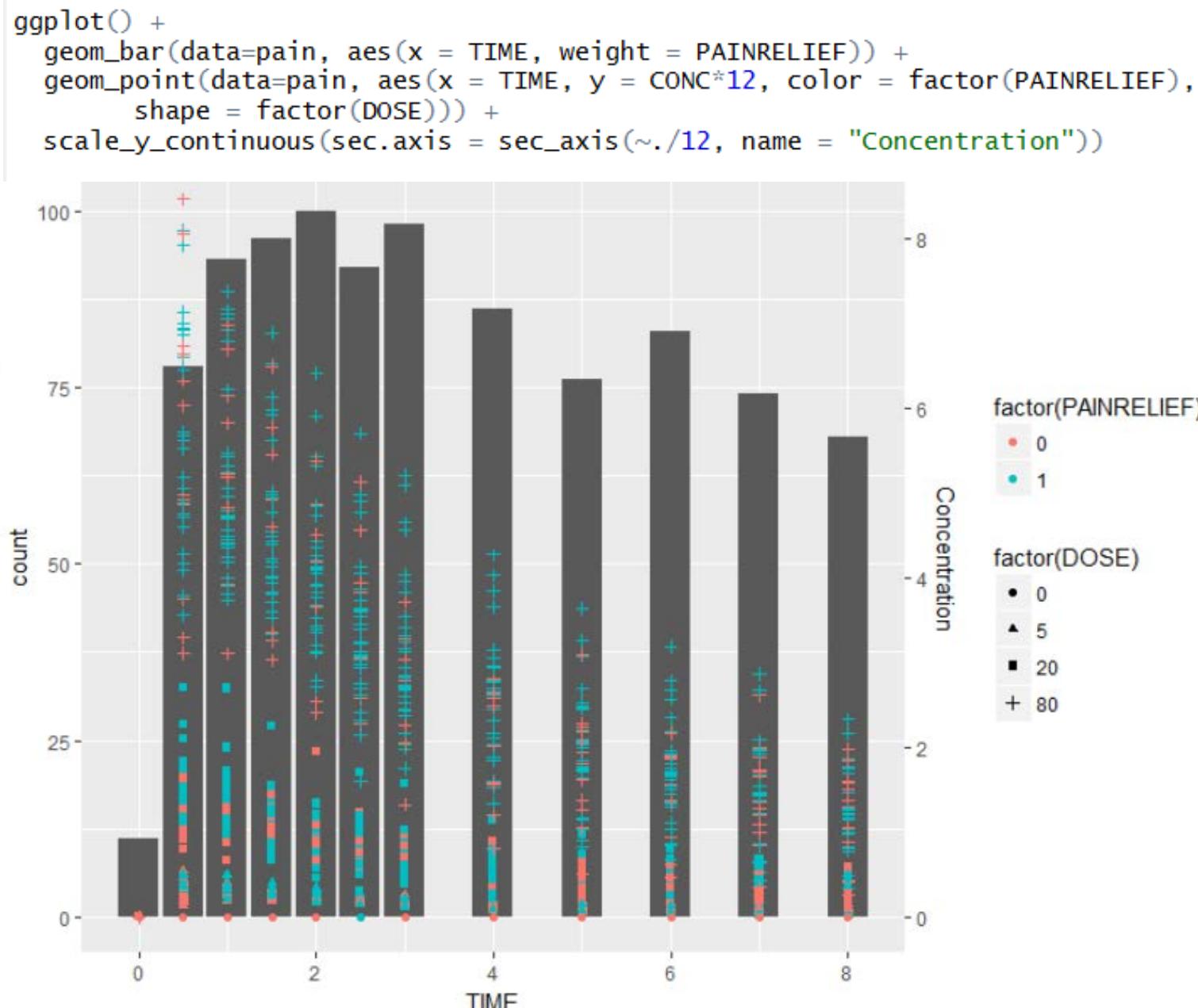
##### Stacked bar chart-----
ggplot(mydf3, aes(x= Dose, y= Count, fill=PainRelief)) +
  geom_bar(stat="identity")

##### count of pain relief by time #####
count4<-ftable(xtabs(~PAINRELIEF+TIME,data=nopain))
mydf4 = melt(count4)
names(mydf4) = c("PainRelief", "Time", "Count")
ggplot(mydf4, aes(x= Time, y= Count, fill=PainRelief)) +
  geom_bar(stat="identity")
```



Preliminary Plots – Exposure and Response vs Time

- Overlaid drug concentration vs time (exposure) with Outcome vs time (response)
- Bars are counts of pain relief at each time point
 - Left y-axis
- Points are drug concentrations
 - Shape of point indicates dose level
 - Color of point indicates whether pain relief occurred at that time



Dose Selection

- Exposure data is continuous
- Response data is discrete (binomial), longitudinal count data
- Logistic regression not appropriate
 - Single instance of binary response
- Proportional odds model (using cumulative logits) not appropriate
 - Single instance of ordinal response
- Poisson regression not appropriate
 - Single instance of count response
- Need marginal models (using GEE) or GLMM (using MLE)

Dose Selection – Marginal Model

- Population-average model using GEE
- Have 5 available predictor variables:
 1. 5 mg Dose
 2. 20 mg Dose
 3. 80 mg Dose
 4. Time point
 5. Drug concentration
- Intercept represents no drug effect (i.e. placebo)

Dose Selection – Marginal Model

- Execute full model first to identify optimal correlation matrix structure
- Then run base model (intercept-only) to assess placebo effect

```
##### Fit a marginal model-----
library(gee)
library(geepack)
pain_gee1 <- gee(PAINRELIEF ~ as.factor(DOSE)+CONC+TIME,
  data = pain, family = "binomial", id = ID,
  corstr = "independence", scale.fix = TRUE, scale.value = 1)
summary(pain_gee1)
```

```
Call:
gee(formula = PAINRELIEF ~ as.factor(DOSE) + CONC + TIME, id = ID,
  data = pain, family = "binomial", corstr = "independence",
  scale.fix = TRUE, scale.value = 1)

Summary of Residuals:
      Min        1Q     Median       3Q      Max 
-0.9680894 -0.4288146 -0.2344756  0.4113285  0.7655244 

Coefficients:
                                         Estimate Naive S.E.   Naive z Robust S.E.   Robust z
(Intercept)                   -1.1832093 0.12443077 -9.508976 0.23110147 -5.1198692
as.factor(DOSE)5               0.6318822 0.13758186  4.592773 0.30106200  2.0988441
as.factor(DOSE)20              0.9842347 0.14450466  6.811094 0.28811949  3.4160644
as.factor(DOSE)80              0.2742298 0.21564776  1.271656 0.33440822  0.8200451
CONC                         0.5057437 0.06511385  7.767068 0.09836741  5.1413741
TIME                          0.0734607 0.02009136  3.656332 0.01755744  4.1840198

Estimated Scale Parameter: 1
Number of Iterations: 1

Working Correlation
      [,1]  [,2]  [,3]  [,4]  [,5]  [,6]  [,7]  [,8]  [,9]  [,10]  [,11]  [,12]
[1,]    1  0  0  0  0  0  0  0  0  0  0  0  0
[2,]    0  1  0  0  0  0  0  0  0  0  0  0  0
[3,]    0  0  1  0  0  0  0  0  0  0  0  0  0
[4,]    0  0  0  1  0  0  0  0  0  0  0  0  0
[5,]    0  0  0  0  1  0  0  0  0  0  0  0  0
[6,]    0  0  0  0  0  1  0  0  0  0  0  0  0
[7,]    0  0  0  0  0  0  1  0  0  0  0  0  0
[8,]    0  0  0  0  0  0  0  1  0  0  0  0  0
[9,]    0  0  0  0  0  0  0  0  1  0  0  0  0
[10,]   0  0  0  0  0  0  0  0  0  1  0  0  0
[11,]   0  0  0  0  0  0  0  0  0  0  1  0  0
[12,]   0  0  0  0  0  0  0  0  0  0  0  0  1
```

Dose Selection – Marginal Model

- Robust SE estimates are not similar to Naïve SE estimates
- “Independence” working correlation is inadequate to assess correlations
- Try another structure

```
Call:  
gee(formula = PAINRELIEF ~ as.factor(DOSE) + CONC + TIME, id = ID,  
    data = pain, family = "binomial", corstr = "independence",  
    scale.fix = TRUE, scale.value = 1)  
  
Summary of Residuals:  
      Min        1Q     Median       3Q      Max  
-0.9680894 -0.4288146 -0.2344756  0.4113285  0.7655244  
  
Coefficients:  
              Estimate  Naive S.E.  Naive z  Robust S.E.  Robust z  
(Intercept) -1.1832093 0.12443077 -9.508976 0.23110147 -5.1198692  
as.factor(DOSE)5 0.6318822 0.13758186  4.592773 0.30106200  2.0988441  
as.factor(DOSE)20 0.9842347 0.14450466  6.811094 0.28811949  3.4160644  
as.factor(DOSE)80 0.2742298 0.21564776  1.271656 0.33440822  0.8200451  
CONC          0.5057437 0.06511385  7.767068 0.09836741  5.1413741  
TIME          0.0734607 0.02009136  3.656332 0.01755744  4.1840198  
  
Estimated Scale Parameter: 1  
Number of Iterations: 1  
  
Working Correlation  
      [,1] [,2] [,3] [,4] [,5] [,6] [,7] [,8] [,9] [,10] [,11] [,12]  
[1,]   1   0   0   0   0   0   0   0   0   0   0   0   0  
[2,]   0   1   0   0   0   0   0   0   0   0   0   0   0  
[3,]   0   0   1   0   0   0   0   0   0   0   0   0   0  
[4,]   0   0   0   1   0   0   0   0   0   0   0   0   0  
[5,]   0   0   0   0   1   0   0   0   0   0   0   0   0  
[6,]   0   0   0   0   0   1   0   0   0   0   0   0   0  
[7,]   0   0   0   0   0   0   1   0   0   0   0   0   0  
[8,]   0   0   0   0   0   0   0   1   0   0   0   0   0  
[9,]   0   0   0   0   0   0   0   0   1   0   0   0   0  
[10,]  0   0   0   0   0   0   0   0   0   1   0   0   0  
[11,]  0   0   0   0   0   0   0   0   0   0   1   0   0  
[12,]  0   0   0   0   0   0   0   0   0   0   0   1   0
```

Dose Selection – Marginal Model

- Tried “exchangeable” matrix structure
- More similarity between Naïve and Robust SE estimates
- Except for CONC variable
- Try another structure

```
##### Exchangeable correlation -----
pain_gee2<- gee(PAINRELIEF ~ as.factor(DOSE)+CONC+TIME, data = pain,
  family = "binomial", id = ID, corstr = "exchangeable",
  scale.fix = TRUE, scale.value = 1)
summary(pain_gee2)
round(summary(pain_gee2)$coefficients,4)
round(summary(pain_gee2)$working.correlation,3)
```

```
> round(summary(pain_gee2)$coefficients,4)
      Estimate Naive S.E. Naive z Robust S.E. Robust z
(Intercept) -1.2213 0.2245 -5.4413 0.2341 -5.2170
as.factor(DOSE)5 0.6680 0.2879 2.3202 0.3033 2.2026
as.factor(DOSE)20 0.9778 0.2925 3.3428 0.2895 3.3780
as.factor(DOSE)80 0.7425 0.3129 2.3732 0.3474 2.1374
CONC          0.5470 0.0667 8.2067 0.1063 5.1451
TIME           0.0744 0.0168 4.4304 0.0170 4.3716
> round(summary(pain_gee2)$working.correlation,3)
 [,1] [,2] [,3] [,4] [,5] [,6] [,7] [,8] [,9] [,10] [,11] [,12]
[1,] 1.000 0.307 0.307 0.307 0.307 0.307 0.307 0.307 0.307 0.307 0.307 0.307
[2,] 0.307 1.000 0.307 0.307 0.307 0.307 0.307 0.307 0.307 0.307 0.307 0.307
[3,] 0.307 0.307 1.000 0.307 0.307 0.307 0.307 0.307 0.307 0.307 0.307 0.307
[4,] 0.307 0.307 0.307 1.000 0.307 0.307 0.307 0.307 0.307 0.307 0.307 0.307
[5,] 0.307 0.307 0.307 0.307 1.000 0.307 0.307 0.307 0.307 0.307 0.307 0.307
[6,] 0.307 0.307 0.307 0.307 0.307 1.000 0.307 0.307 0.307 0.307 0.307 0.307
[7,] 0.307 0.307 0.307 0.307 0.307 0.307 1.000 0.307 0.307 0.307 0.307 0.307
[8,] 0.307 0.307 0.307 0.307 0.307 0.307 0.307 1.000 0.307 0.307 0.307 0.307
[9,] 0.307 0.307 0.307 0.307 0.307 0.307 0.307 0.307 1.000 0.307 0.307 0.307
[10,] 0.307 0.307 0.307 0.307 0.307 0.307 0.307 0.307 0.307 1.000 0.307 0.307
[11,] 0.307 0.307 0.307 0.307 0.307 0.307 0.307 0.307 0.307 0.307 1.000 0.307
[12,] 0.307 0.307 0.307 0.307 0.307 0.307 0.307 0.307 0.307 0.307 0.307 1.000
```

Dose Selection – Marginal Model

- Tried “auto-regressive” matrix structure
- Decent similarity between Naïve and Robust SE estimates
- Try another structure anyway

```
#### Auto-regressive correl matix---  
pain_gee3<-gee(PAINRELIEF~ as.factor(DOSE)+CONC+TIME, data = pain,  
  family="binomial", id=ID, corstr="AR-M",  
  scale.fix=TRUE, scale.value=1)  
summary(pain_gee3)  
round(summary(pain_gee3)$coefficients,4)  
round(summary(pain_gee3)$working.correlation,3)
```

```
> round(summary(pain_gee3)$coefficients,4)  
             Estimate Naive S.E. Naive z Robust S.E. Robust z  
(Intercept) -1.2982   0.1778 -7.3016   0.2295 -5.6569  
as.factor(DOSE)5  0.5932   0.2027  2.9257   0.2996  1.9799  
as.factor(DOSE)20  0.9068   0.2070  4.3807   0.2849  3.1827  
as.factor(DOSE)80  0.1953   0.2595  0.7528   0.3278  0.5959  
CONC            0.5787   0.0713  8.1160   0.1161  4.9866  
TIME             0.0861   0.0254  3.3931   0.0179  4.8244  
> round(summary(pain_gee3)$working.correlation,3)  
      [,1]  [,2]  [,3]  [,4]  [,5]  [,6]  [,7]  [,8]  [,9]  [,10]  [,11]  [,12]  
[1,] 1.000 0.406 0.164 0.067 0.027 0.011 0.004 0.002 0.001 0.000 0.000 0.000  
[2,] 0.406 1.000 0.406 0.164 0.067 0.027 0.011 0.004 0.002 0.001 0.000 0.000  
[3,] 0.164 0.406 1.000 0.406 0.164 0.067 0.027 0.011 0.004 0.002 0.001 0.000  
[4,] 0.067 0.164 0.406 1.000 0.406 0.164 0.067 0.027 0.011 0.004 0.002 0.001  
[5,] 0.027 0.067 0.164 0.406 1.000 0.406 0.164 0.067 0.027 0.011 0.004 0.002  
[6,] 0.011 0.027 0.067 0.164 0.406 1.000 0.406 0.164 0.067 0.027 0.011 0.004  
[7,] 0.004 0.011 0.027 0.067 0.164 0.406 1.000 0.406 0.164 0.067 0.027 0.011  
[8,] 0.002 0.004 0.011 0.027 0.067 0.164 0.406 1.000 0.406 0.164 0.067 0.027  
[9,] 0.001 0.002 0.004 0.011 0.027 0.067 0.164 0.406 1.000 0.406 0.164 0.067  
[10,] 0.000 0.001 0.002 0.004 0.011 0.027 0.067 0.164 0.406 1.000 0.406 0.164  
[11,] 0.000 0.000 0.001 0.002 0.004 0.011 0.027 0.067 0.164 0.406 1.000 0.406  
[12,] 0.000 0.000 0.000 0.001 0.002 0.004 0.011 0.027 0.067 0.164 0.406 1.000  
> |
```

Dose Selection – Marginal Model

- Tried “unstructured” matrix structure
- Good similarity between Naïve and Robust SE estimates
- Unstructured is the best choice

```
#### Unstructured matrix--  
pain_gee4<-gee(PAINRELIEF~as.factor(DOSE)+CONC+TIME, data=pain,  
family="binomial", id=ID, corstr="unstructured",  
scale.fix=TRUE, scale.value=1)  
summary(pain_gee4)  
round(summary(pain_gee4)$coefficients,4)  
round(summary(pain_gee4)$working.correlation,3)  
####Looks the best----
```

```
> round(summary(pain_gee4)$coefficients,4)  
          Estimate Naive S.E. Naive z Robust S.E. Robust z  
(Intercept) -1.4464   0.2196 -6.5860   0.2253 -6.4213  
as.factor(DOSE)5  0.6113   0.2799  2.1840   0.2925  2.0902  
as.factor(DOSE)20  1.0676   0.2759  3.8687   0.2660  4.0128  
as.factor(DOSE)80  0.7757   0.2938  2.6404   0.2851  2.7210  
CONC           0.3120   0.0591  5.2783   0.0636  4.9035  
TIME            0.0635   0.0177  3.5849   0.0168  3.7809  
> round(summary(pain_gee4)$working.correlation,3)  
      [,1]  [,2]  [,3]  [,4]  [,5]  [,6]  [,7]  [,8]  [,9]  [,10]  
[1,]  1.000  0.074 -0.015 -0.037 -0.076 -0.081 -0.113 -0.017  0.118  0.020  
[2,]  0.074  1.000  0.501  0.416  0.373  0.366  0.499  0.311  0.354  0.458  
[3,] -0.015  0.501  1.000  0.474  0.464  0.447  0.519  0.374  0.360  0.411  
[4,] -0.037  0.416  0.474  1.000  0.588  0.551  0.487  0.399  0.535  0.433  
[5,] -0.076  0.373  0.464  0.588  1.000  0.570  0.512  0.336  0.357  0.371  
[6,] -0.081  0.366  0.447  0.551  0.570  1.000  0.440  0.367  0.353  0.471  
[7,] -0.113  0.499  0.519  0.487  0.512  0.440  1.000  0.430  0.370  0.475  
[8,] -0.017  0.311  0.374  0.399  0.336  0.367  0.430  1.000  0.362  0.342  
[9,]  0.118  0.354  0.360  0.535  0.357  0.353  0.370  0.362  1.000  0.468  
[10,] 0.020  0.458  0.411  0.433  0.371  0.471  0.475  0.342  0.468  1.000  
[11,] 0.109  0.382  0.295  0.206  0.346  0.315  0.324  0.240  0.292  0.296  
[12,] 0.153  0.237  0.295  0.267  0.439  0.280  0.204  0.270  0.430  0.294
```

Marginal Model – Intercept Only

- Using unstructured correlation matrix, assess baseline effect (prior to drug/placebo treatment)

```
#### Baseline model; intercept-only model---
pain_geeBL<-gee(PAINRELIEF~1, data=pain, family="binomial",
                   id=ID, corstr="unstructured", scale.fix=TRUE, scale.value=1)
summary(pain_geeBL)
round(summary(pain_geeBL)$coefficients,4)
round(summary(pain_geeBL)$working.correlation,3)
```

```
--, -----
> round(summary(pain_geeBL)$coefficients,4)
      Estimate Naive S.E. Naive z Robust S.E. Robust z
(Intercept) -0.5403    0.0904 -5.9735     0.0874 -6.1837
> round(summary(pain_geeBL)$working.correlation,3)
      [,1]   [,2]   [,3]   [,4]   [,5]   [,6]   [,7]   [,8]   [,9]   [,10]  [,11]  [,12]
[1,] 1.000 -0.049 -0.155 -0.157 -0.192 -0.170 -0.198 -0.094  0.015 -0.068  0.009  0.038
[2,] -0.049  1.000  0.533  0.507  0.472  0.471  0.583  0.429  0.445  0.572  0.462  0.327
[3,] -0.155  0.533  1.000  0.565  0.554  0.529  0.618  0.487  0.456  0.536  0.380  0.385
[4,] -0.157  0.507  0.565  1.000  0.692  0.667  0.615  0.531  0.618  0.557  0.354  0.382
[5,] -0.192  0.472  0.554  0.692  1.000  0.680  0.628  0.474  0.466  0.523  0.460  0.535
[6,] -0.170  0.471  0.529  0.667  0.680  1.000  0.580  0.519  0.465  0.592  0.435  0.393
[7,] -0.198  0.583  0.618  0.615  0.628  0.580  1.000  0.561  0.483  0.610  0.454  0.341
[8,] -0.094  0.429  0.487  0.531  0.474  0.519  0.561  1.000  0.470  0.480  0.370  0.375
[9,]  0.015  0.445  0.456  0.618  0.466  0.465  0.483  0.470  1.000  0.566  0.386  0.484
[10,] -0.068  0.572  0.536  0.557  0.523  0.592  0.610  0.480  0.566  1.000  0.419  0.401
[11,]  0.009  0.462  0.380  0.354  0.460  0.435  0.454  0.370  0.386  0.419  1.000  0.455
[12,]  0.038  0.327  0.385  0.382  0.535  0.393  0.341  0.375  0.484  0.401  0.455  1.000
```

Marginal Model – Intercept Only

- Assess statistical significance of intercept as a predictive variable

```
#### calculate p-value ----  
pval.BL<-pnorm(abs(teststat.geeBL[,5]),lower.tail=FALSE)*2  
## final table-----  
cbind(teststat.geeBL,pval.BL)
```

	Estimate	Naive S.E.	Naive z	Robust S.E.	Robust z	pval.BL
(Intercept)	-0.54025	0.09044	-5.97351	0.08737	-6.18366	6.263219e-10

- Exponentiate coefficient to get odds ratio (w/ 95% CI)

```
##### odds ratio and CI for baseline as predictor---  
se.BL <- summary(pain_geeBL)$coefficients["(Intercept)","Robust S.E."]  
ci.BL<-exp(coef(pain_geeBL)["(Intercept)"] + c(-1,1) * se.BL * qnorm(0.975))  
cbind(OR=exp(coef(pain_geeBL)["(Intercept)"]),ci.BL)
```

	OR	ci.BL
[1,]	0.5826018	0.4909129
[2,]	0.5826018	0.6914156

Inference: the baseline odds of a typical person not on study or prior to study of having pain relief is between 0.49x – 0.69x that of a typical person on the placebo arm.

Marginal Model – Full

- “unstructured” matrix structure
 - Same as a few slides ago
- Get p-values...

```
#### Unstructured matrix---  
pain_gee4<-gee(PAINRELIEF~as.factor(DOSE)+CONC+TIME, data=pain,  
family="binomial", id=ID, corstr="unstructured",  
scale.fix=TRUE, scale.value=1)  
summary(pain_gee4)  
round(summary(pain_gee4)$coefficients,4)  
round(summary(pain_gee4)$working.correlation,3)  
####Looks the best----
```

```
> round(summary(pain_gee4)$coefficients,4)  
          Estimate Naive S.E. Naive z Robust S.E. Robust z  
(Intercept) -1.4464   0.2196 -6.5860   0.2253 -6.4213  
as.factor(DOSE)5  0.6113   0.2799  2.1840   0.2925  2.0902  
as.factor(DOSE)20  1.0676   0.2759  3.8687   0.2660  4.0128  
as.factor(DOSE)80  0.7757   0.2938  2.6404   0.2851  2.7210  
CONC           0.3120   0.0591  5.2783   0.0636  4.9035  
TIME            0.0635   0.0177  3.5849   0.0168  3.7809  
> round(summary(pain_gee4)$working.correlation,3)  
      [,1]  [,2]  [,3]  [,4]  [,5]  [,6]  [,7]  [,8]  [,9]  [,10]  
[1,]  1.000  0.074 -0.015 -0.037 -0.076 -0.081 -0.113 -0.017  0.118  0.020  
[2,]  0.074  1.000  0.501  0.416  0.373  0.366  0.499  0.311  0.354  0.458  
[3,] -0.015  0.501  1.000  0.474  0.464  0.447  0.519  0.374  0.360  0.411  
[4,] -0.037  0.416  0.474  1.000  0.588  0.551  0.487  0.399  0.535  0.433  
[5,] -0.076  0.373  0.464  0.588  1.000  0.570  0.512  0.336  0.357  0.371  
[6,] -0.081  0.366  0.447  0.551  0.570  1.000  0.440  0.367  0.353  0.471  
[7,] -0.113  0.499  0.519  0.487  0.512  0.440  1.000  0.430  0.370  0.475  
[8,] -0.017  0.311  0.374  0.399  0.336  0.367  0.430  1.000  0.362  0.342  
[9,]  0.118  0.354  0.360  0.535  0.357  0.353  0.370  0.362  1.000  0.468  
[10,] 0.020  0.458  0.411  0.433  0.371  0.471  0.475  0.342  0.468  1.000  
[11,] 0.109  0.382  0.295  0.206  0.346  0.315  0.324  0.240  0.292  0.296  
[12,] 0.153  0.237  0.295  0.267  0.439  0.280  0.204  0.270  0.430  0.294
```

Marginal Model – Full

- All variables significantly predict response outcome
- Intercept and Conc has p=0.0?
- Both highly significant
- Full model =
Final model

```
##### Full Model Hypothesis Testing-----
teststat.gee4<-round(summary(pain_gee4)$coefficients,4)
#### calculate p-value -----
pval4<-round(pnorm(abs(teststat.gee4[,5]),lower.tail=FALSE)^2,4)
## final table-----
cbind(teststat.gee4,pval4)
```

	Estimate	Naive S.E.	Naive z	Robust S.E.	Robust z	pval4
(Intercept)	-1.4464	0.2196	-6.5860	0.2253	-6.4213	0.0000
as.factor(DOSE)5	0.6113	0.2799	2.1840	0.2925	2.0902	0.0366
as.factor(DOSE)20	1.0676	0.2759	3.8687	0.2660	4.0128	0.0001
as.factor(DOSE)80	0.7757	0.2938	2.6404	0.2851	2.7210	0.0065
CONC	0.3120	0.0591	5.2783	0.0636	4.9035	0.0000
TIME	0.0635	0.0177	3.5849	0.0168	3.7809	0.0002
>						

	Estimate	Naive S.E.	Naive z	Robust S.E.	Robust z	pval4
(Intercept)	-1.44641	0.21962	-6.58595	0.22525	-6.42131	1.351066e-10
as.factor(DOSE)5	0.61128	0.27989	2.18396	0.29245	2.09018	3.660163e-02
as.factor(DOSE)20	1.06756	0.27595	3.86871	0.26604	4.01281	6.000021e-05
as.factor(DOSE)80	0.77570	0.29379	2.64035	0.28508	2.72104	6.507689e-03
CONC	0.31203	0.05912	5.27831	0.06363	4.90352	9.413435e-07
TIME	0.06354	0.01772	3.58492	0.01680	3.78094	1.562373e-04

Dose Response Relationship

- Exponentiate to get OR (w/ 95% CI) for each coefficient

0 mg dose

5 mg dose

20 mg dose

80 mg dose

OR	ci.BL
0.5826018	0.4909129
0.5826018	0.6914156

OR	ci.dose5
1.842785	1.038817
1.842785	3.268963

OR	ci.dose20
2.675763	1.521248
2.675763	4.706472

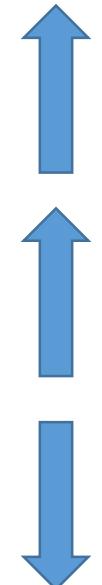
OR	ci.dose80
1.315517	0.6830425
1.315517	2.5336423

```
### interpretation of population averages through odds ratios---  
##### Oddsratio and confidence interval-----  
se.dose5 <- summary(pain_gee4)$coefficients["as.factor(DOSE)5","Robust S.E."]  
ci.dose5<-exp(coef(pain_gee4)["as.factor(DOSE)5"] + c(-1,1) * se.dose5 * qnorm(0.975))  
cbind(OR=exp(coef(pain_gee4)["as.factor(DOSE)5"]),ci.dose5)  
  
se.dose20<- summary(pain_gee1)$coefficients["as.factor(DOSE)20","Robust S.E."]  
ci.dose20<-exp(coef(pain_gee1)["as.factor(DOSE)20"] + c(-1,1) * se.dose20 * qnorm(0.975))  
cbind(OR=exp(coef(pain_gee1)["as.factor(DOSE)20"]),ci.dose20)  
  
se.dose80<- summary(pain_gee1)$coefficients["as.factor(DOSE)80","Robust S.E."]  
ci.dose80<-exp(coef(pain_gee1)["as.factor(DOSE)80"] + c(-1,1) * se.dose80 * qnorm(0.975))  
cbind(OR=exp(coef(pain_gee1)["as.factor(DOSE)80"]),ci.dose80)
```

Dose Response Relationship

<u>0 mg dose</u>	<u>5 mg dose</u>	<u>20 mg dose</u>	<u>80 mg dose</u>
OR ci.BL 0.5826018 0.4909129	OR ci.dose5 1.842785 1.038817	OR ci.dose20 2.675763 1.521248	OR ci.dose80 1.315517 0.6830425
0.5826018 0.6914156	1.842785 3.268963	2.675763 4.706472	1.315517 2.5336423

- “typical person” odds of having pain relief is $0.49x - 0.69x$ that of a “typical person” taking placebo (0 mg)
- “typical person” odds of having pain relief is $1.04x - 3.27x$ that of a “typical person” taking placebo (5 mg)
- “typical person” odds of having pain relief is $1.52x - 4.71x$ that of a “typical person” taking placebo (20 mg)
- “typical person” odds of having pain relief is $0.68x - 2.54x$ that of a “typical person” taking placebo (80 mg)



Dose Response Relationship

- “typical person” odds of having pain relief is $0.49x - 0.69x$ that of a “typical person” taking placebo (0 mg)
 - “typical person” odds of having pain relief is $1.04x - 3.27x$ that of a “typical person” taking placebo (5 mg)
 - “typical person” odds of having pain relief is $1.52x - 4.71x$ that of a “typical person” taking placebo (20 mg)
 - “typical person” odds of having pain relief is $0.68x - 2.54x$ that of a “typical person” taking placebo (80 mg)
- 
- No increased benefit at 80 mg relative to 20 mg, even with increased AUC from 20mg to 80mg

Dose – Exposure Relationship

Perform NCA to obtain Cmax, AUC exposure metrics per subject

```
#####
##### NCA analysis on pain data to get PK values #####
install.packages("PKNCA")
### Perform NCA #####
library(PKNCA)
library(knitr)

## By default it is groupedData; convert it to a data frame for use--
### maps conc to any variable or grouping factor ##
my.conc<-PKNCAdconc(as.data.frame(pain), CONC~TIME|ID)

## Dosing data needs to only have one row per dose, so subset for that first##
d.dose<-unique(pain[pain$TIME == 0.0,
                     c("ARM","ID","TIME","CONC", "PAINRELIEF","DOSE")])

knitr::kable(d.dose,
             caption="Dosing data extracted from data set")

### by dose
my.dose <- PKNCAdose(d.dose, DOSE~TIME|ID)
my.dose
#### combine dose and conc data ##
my.data.automatic <- PKNCAdata(my.conc, my.dose)
summary(my.data.automatic)
knitr::kable(PKNCA.options("single.dose.aucs"))
knitr::kable(my.data.automatic$intervals)
```

```
### specify start/stop times #####
my.intervals <- data.frame(start=0.0,
                            end=Inf,
                            cmax=TRUE,
                            tmax=TRUE,
                            aucinf=TRUE,
                            auclast=TRUE,
                            cl=TRUE,
                            vz=TRUE,
                            half.life=TRUE)

my.data.manual <- PKNCAdata(my.conc, my.dose,intervals=my.intervals)
knitr::kable(my.data.manual$interval)
summary(my.data.manual)

my.results.manual <- pk.nca(my.data.manual)
summary(my.results.manual)
my.results.manual$result

write.csv(my.results.manual$result, file="PKNCA data.csv")
knitr::kable(summary(my.results.manual))
### make a data.frame for PK manual results###
PKresults<-my.results.manual$result
#### drop first two columns in PK results ####
PKresults1<-PKresults[c(-1, -2)]
#### subset out only t=0hr rows in main data.frame ####
pain.0<-pain[pain$TIME==0,]

PKresults2<-t(PKresults1)
### melt PK data ####
mPKresults<-melt(PKresults1, id=c("ID", "PPTESTCD"))
### cast the melted data ##
### cast(data, formula, function) ####
PKresults3<-dcast(mPKresults, ID~variable+PPTESTCD)
write.csv(PKresults3, file="PKNCA results.csv")
#### Cbind pk results to subsetted data.frame ####
pain.nca<-cbind2(PKresults3, pain.0)
head(pain.nca)
write.csv(pain.nca, file="FinalPKNCA.csv")
### drop columns except cmax, auclast, HL ####
pain.nca1<-pain.nca[c(-2, -4,-7,-8,-9,-10,-11,-12,-13,-14,-15,-16,-17,-18,
                     -19,-20,-21,-22,-23,-24,-25,-27)]
write.csv(pain.nca1, file="FinalPKNCA1.csv")
```

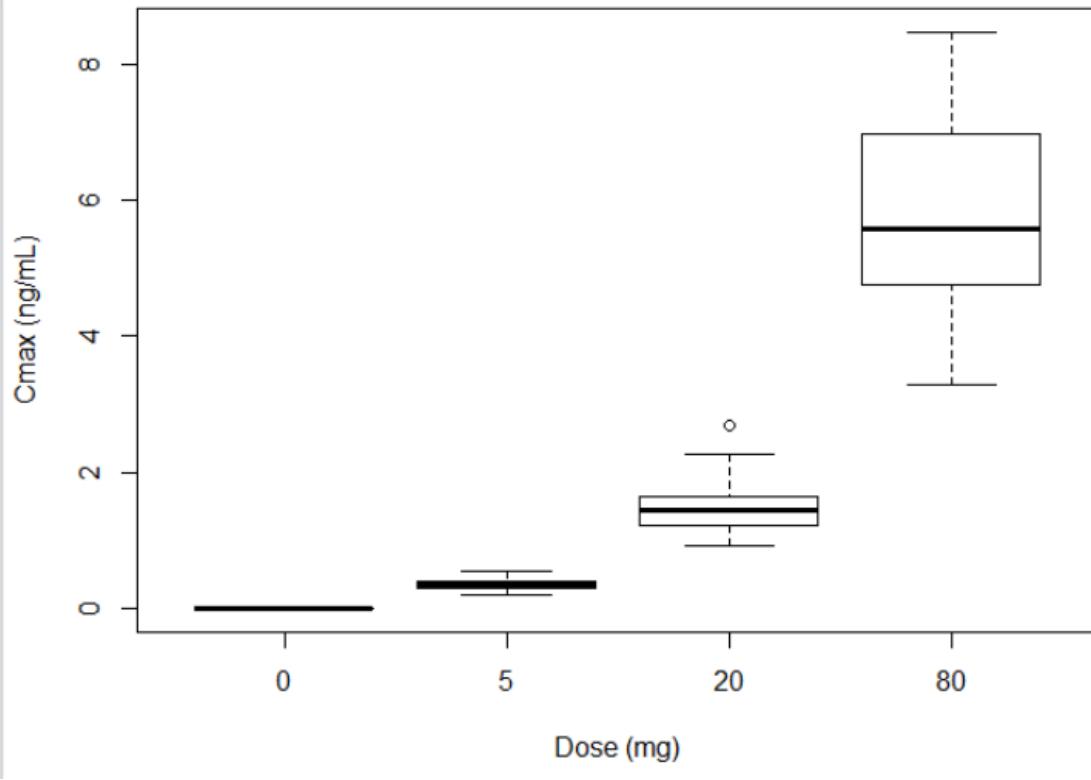
NCA Results

```
nca<-read.csv("FinalPKNCA1.csv", header=TRUE)

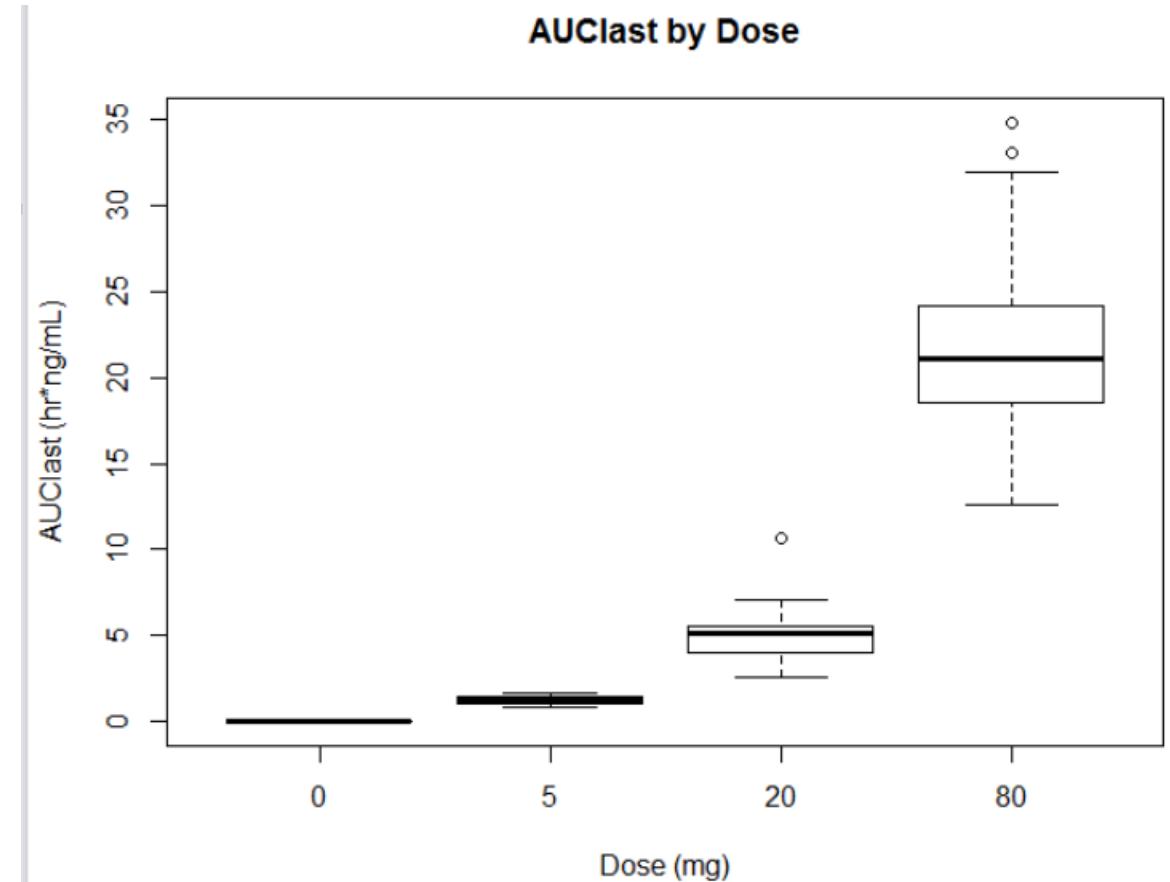
### plot overlay of dose vs AUC, with dose vs PainRelief ####
## to show that increasing exposure does NOT increase response ####
boxplot(PPORRES_auclast ~ DOSE, data = nca, main="AUClast by Dose", xlab="Dose (mg)",
       ylab="AUClast (hr*ng/mL)")

boxplot(PPORRES_cmax ~ DOSE, data = nca, main="Cmax by Dose", xlab="Dose (mg)",
       ylab="Cmax (ng/mL)")
```

Cmax by Dose



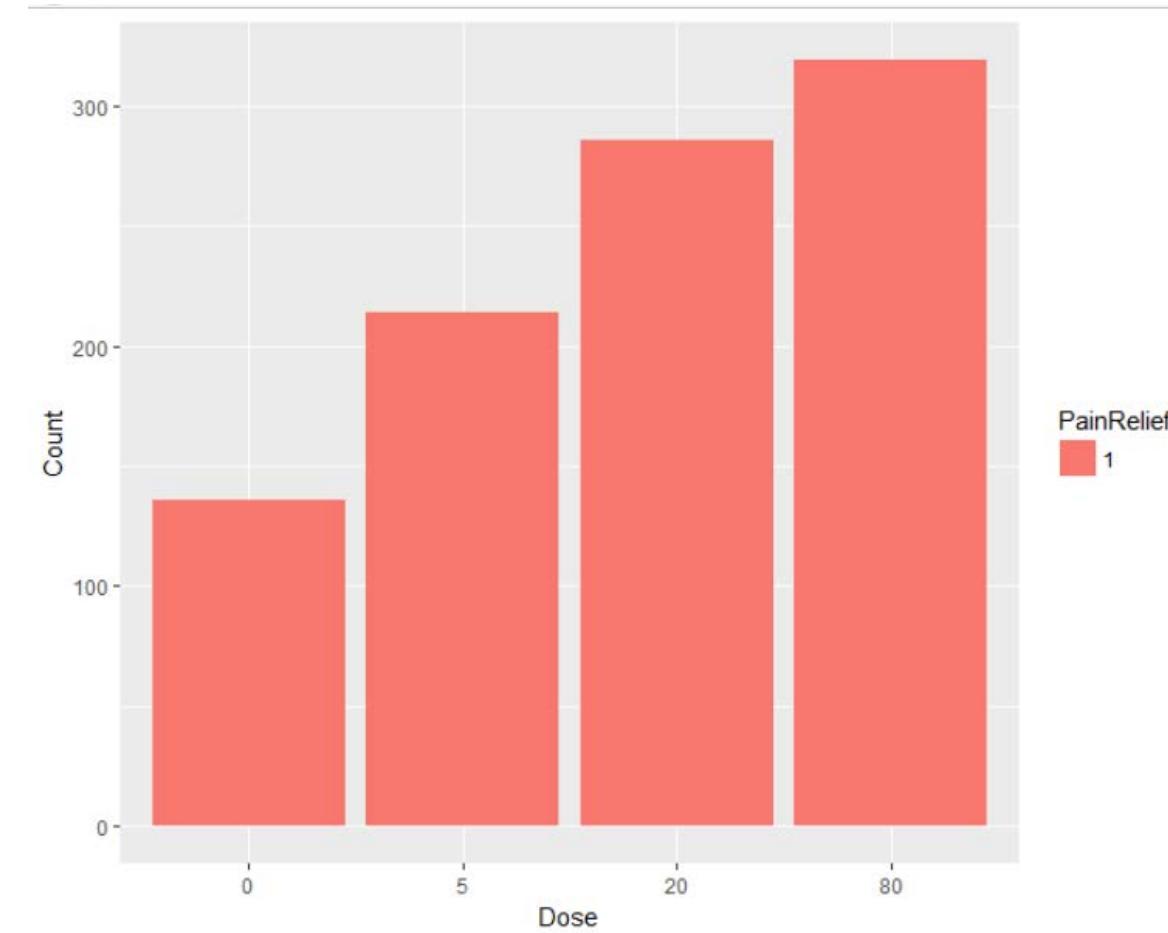
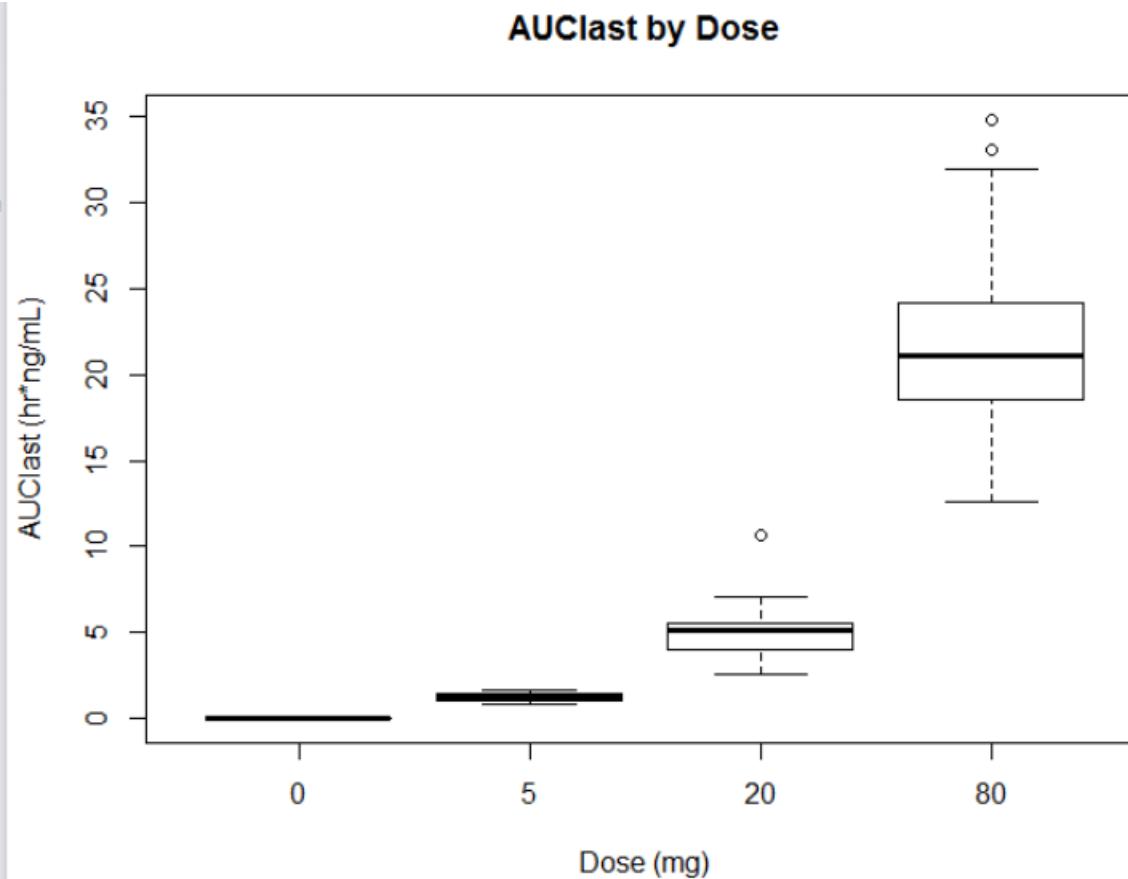
AUClast by Dose



Dose – Response Relationship

Dose increases exposure linearly. But Response tails off with Dose

```
plot6<-plot(PPORRES_auctlast~DOSE, data=pain.nca1, main="AUC by Dose",
             xlab="Dose(mg)", ylab="AUC(hr*ng/mL)",
             cex=1.8,cex.lab=1.5,cex.axis=1.2, col="black")
```



Dose – Response Relationship

- Marginal model via GEE suggested 20mg dose is optimal
 - No added benefit with 80mg
 - Odds ratio for typical person experiencing pain relief at 80mg is LOWER than odds ratio at 20 mg
- That is supported by observational plots of data
 - PK exposure (Cmax and AUC) increases linearly with dose (dose proportional)
 - Response increases less than dose proportionally from 20mg to 80mg
- Both analyses suggest no added therapeutic benefit from 80mg

Dose Selection - GLMM

- Use GLMM to obtain subject-specific mean responses based on predictor variables
 - In this case, binary (yes/no) to pain relief
- Initially, try intercept-only (Base model) for baseline effect (prior to therapy)

GLMM – Base Model

- Baseline (pre-dose) was NOT a significant predictor of pain relief

```
#### Generalized linear mixed effects model (GLMM)---  
library(lme4)
```

```
#### Baseline model (intercept-only)---  
pain.glmmBL<-glmer(PAINRELIEF ~ 1+(1|ID),data = pain, family = binomial, REML=F)  
pain.glmmBL  
summary(pain.glmmBL)  
#### FULL model---  
pain.glmm<-glmer(PAINRELIEF ~ as.factor(DOSE)+TIME+CONC+(1|ID),data = pain, family = binomial, REML=F)  
summary(pain.glmm)  
#####Test the significance of adding ALL covariates = Full model vs base-----  
pain.glmmBL$logLik  
pain.glmm$logLik
```

Family: binomial (logit)
Formula: PAINRELIEF ~ 1 + (1 | ID)
Data: pain

AIC	BIC	logLik	deviance	df.resid
2182.3	2193.4	-1089.1	2178.3	1918

Scaled residuals:

Min	1Q	Median	3Q	Max
-2.5186	-0.5325	-0.2745	0.6206	2.5472

Random effects:

Groups	Name	Variance	Std.Dev.
ID	(Intercept)	3.006	1.734
Number of obs:	1920, groups:	ID, 160	

Fixed effects:

	Estimate	Std. Error	z value	Pr(> z)
(Intercept)	-0.05877	0.14991	-0.392	0.695

GLMM – FULL Model

- Baseline (pre-dose) is now a significant predictor of pain relief
- So too is 5mg and 20mg, sampling time post dose, and drug concentration
- 80 mg is not a significant predictor
- Model diagnostics all improved vs Base model

```
#### FULL model---  
pain.glmm<-glmer(PAINRELIEF ~ as.factor(DOSE)+TIME+CONC+(1|ID),data = pain,  
family = binomial, REML=F)  
summary(pain.glmm)
```

```
Family: binomial ( logit )  
Formula: PAINRELIEF ~ as.factor(DOSE) + TIME + CONC + (1 | ID)  
Data: pain
```

AIC	BIC	LogLik	deviance	df.resid
2055.3	2094.2	-1020.7	2041.3	1913

Scaled residuals:

Min	1Q	Median	3Q	Max
-5.7905	-0.5419	-0.1832	0.5806	3.3807

Random effects:

Groups	Name	Variance	Std.Dev.
ID	(Intercept)	2.574	1.604
Number of obs:	1920, groups:	ID, 160	

Fixed effects:

	Estimate	Std. Error	z value	Pr(> z)
(Intercept)	-1.78397	0.30383	-5.872	4.31e-09 ***
as.factor(DOSE)5	0.96300	0.40336	2.387	0.016965 *
as.factor(DOSE)20	1.43220	0.40558	3.531	0.000414 ***
as.factor(DOSE)80	0.42782	0.45373	0.943	0.345734
TIME	0.10906	0.02417	4.512	6.42e-06 ***
CONC	0.72685	0.08391	8.663	< 2e-16 ***

Signif. codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Correlation of Fixed Effects:

	(Intr)	a.(DOSE)5	a.(DOSE)2	a.(DOSE)8	TIME
as.f(DOSE)5	-0.686				
as.(DOSE)20	-0.676	0.516			
as.(DOSE)80	-0.564	0.469	0.512		
TIME	-0.297	0.005	-0.022	-0.146	
CONC	-0.112	-0.020	-0.116	-0.437	0.324

Hypothesis Testing of Models (Base vs Full)

- Using the “log-likelihood” statistic for each model, use the likelihood ratio test (LRT)
- Base model has 1 variable
- Full model has 6 variables
- Degrees of freedom = 5
- Chi-squared test statistic at alpha=0.05, w/ df=5 = 11.07
- $(-2 \cdot LL_{Full}) - (-2 \cdot LL_{Base})$

```
> lrt.BLFull = -2*(((-1020.7) - (-1089.131))  
> lrt.BLFull  
[1] 136.862  
> qchisq(0.95,5)  
[1] 11.0705
```

```
-----  
install.packages("lmtree")  
library(lmtree)  
lrtest(pain.g1mm, pain.g1mmBL)  
.....  
Likelihood ratio test  
  
Model 1: PAINRELIEF ~ as.factor(DOSE) + TIME + CONC + (1 | ID)  
Model 2: PAINRELIEF ~ 1 + (1 | ID)  
#Df LogLik Df Chisq Pr(>Chisq)  
1 7 -1020.7  
2 2 -1089.1 -5 136.95 < 2.2e-16 ***  
---  
Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Full model is significantly a better fit than Base Model

GLMM – FULL Model

- There could be variability in drug concentrations between patients given the same dose
 - E.g. not every patient given 20mg had the same Cmax and AUC
- To account for this, add a random effect parameter to the “CONC” variable, in addition to “ID”

```
### Added random effect to CONC and ID--
pain.glmm1<-glmer(PAINRELIEF ~ as.factor(DOSE)+TIME+CONC+(1|ID)+(1|CONC),
                     data = pain, family = binomial, REML=TRUE)
summary(pain.glmm1)
lrtest(pain.glmm1, pain.glmm)

Family: binomial ( logit )
Formula: PAINRELIEF ~ as.factor(DOSE) + TIME + CONC + (1 | ID) + (1 | CONC)
Data: pain

AIC      BIC      LogLik deviance df.resid
1960.8  2005.3   -972.4    1944.8     1912

Scaled residuals:
    Min      1Q  Median      3Q     Max 
-2.80807 -0.47784 -0.05215  0.49314  3.01931 

Random effects:
Groups  Name        Variance Std.Dev. 
CONC   (Intercept) 0.8082   0.899  
ID     (Intercept) 3.6672   1.915  
Number of obs: 1920, groups: CONC, 1319; ID, 160

Fixed effects:
            Estimate Std. Error z value Pr(>|z|)    
(Intercept)  2.45984   0.62302  3.948 7.87e-05 ***
as.factor(DOSE)5 -2.48540   0.64791 -3.836 0.000125 ***
as.factor(DOSE)20 -1.60201   0.61822 -2.591 0.009561 ** 
as.factor(DOSE)80 -1.64440   0.61144 -2.689 0.007158 ** 
TIME         -0.04231   0.03267 -1.295 0.195272    
CONC          0.30782   0.10102  3.047 0.002311 ** 
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Correlation of Fixed Effects:
              (Intr) a.(DOSE)5 a.(DOSE)2 a.(DOSE)8 TIME
as.f(DOSE)5 -0.838
as.(DOSE)20 -0.810  0.713
as.(DOSE)80 -0.668  0.622    0.643
TIME         -0.505  0.282    0.239    0.065
CONC          -0.362  0.222    0.146    -0.185   0.486
convergence code: 0
Model failed to converge with max|grad| = 0.00107671 (tol = 0.001, component 1)
```

GLMM – FULL Model

- Gives an error message of model failing to converge
- This can be ignored, as long as the value given is <0.01
- In our case, value=0.001076

```
### Added random effect to CONC and ID--  
pain.glmm1<-glmer(PAINRELIEF ~ as.factor(DOSE)+TIME+CONC+(1|ID)+(1|CONC),  
                     data = pain, family = binomial, REML=TRUE)  
summary(pain.glmm1)
```

```
Family: binomial ( logit )  
Formula: PAINRELIEF ~ as.factor(DOSE) + TIME + CONC + (1 | ID) + (1 | CONC)  
Data: pain  
  
AIC      BIC      logLik deviance df.resid  
1960.8   2005.3   -972.4    1944.8     1912  
  
Scaled residuals:  
    Min     1Q   Median     3Q    Max  
-2.80807 -0.47784 -0.05215  0.49314  3.01931  
  
Random effects:  
Groups Name        Variance Std.Dev.  
CONC   (Intercept) 0.8082   0.899  
ID     (Intercept) 3.6672   1.915  
Number of obs: 1920, groups: CONC, 1319; ID, 160  
  
Fixed effects:  
            Estimate Std. Error z value Pr(>|z|)  
(Intercept)  2.45984   0.62302  3.948 7.87e-05 ***  
as.factor(DOSE)5 -2.48540   0.64791 -3.836 0.000125 ***  
as.factor(DOSE)20 -1.60201   0.61822 -2.591 0.009561 **  
as.factor(DOSE)80 -1.64440   0.61144 -2.689 0.007158 **  
TIME          -0.04231   0.03267 -1.295 0.195272  
CONC          0.30782   0.10102  3.047 0.002311 **  
---  
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1  
  
Correlation of Fixed Effects:  
              (Intr) a.(DOSE)5 a.(DOSE)2 a.(DOSE)8 TIME  
as.f(DOSE)5 -0.838  
as.(DOSE)20 -0.810  0.713  
as.(DOSE)80 -0.668  0.622    0.643  
TIME         -0.505  0.282    0.239    0.065  
CONC         -0.362  0.222    0.146   -0.185    0.486  
convergence code: 0  
Model failed to converge with max|grad| = 0.00107671 (tol = 0.001, component 1)
```

GLMM – FULL Model

- Now, all dose levels are significant predictors
- Pre-dose (intercept) is a significant predictor
- Drug concentrations are
- Sampling times post dose are NOT
- Significant improvement from Full model w/o BSV on CONC

```
### Added random effect to CONC and ID--  
pain.glmm1<-glmer(PAINRELIEF ~ as.factor(DOSE)+TIME+CONC+(1|ID)+(1|CONC),  
                     data = pain, family = binomial, REML=TRUE)  
summary(pain.glmm1)
```

```
Family: binomial ( logit )  
Formula: PAINRELIEF ~ as.factor(DOSE) + TIME + CONC + (1 | ID) + (1 | CONC)  
Data: pain  
  
AIC      BIC      LogLik deviance df.resid  
1960.8  2005.3   -972.4    1944.8     1912  
  
Scaled residuals:  
    Min     1Q   Median     3Q    Max  
-2.80807 -0.47784 -0.05215  0.49314  3.01931  
  
Random effects:  
Groups Name        Variance Std.Dev.  
CONC   (Intercept) 0.8082   0.899  
ID     (Intercept) 3.6672   1.915  
Number of obs: 1920, groups: CONC, 1319; ID, 160  
  
Fixed effects:  
            Estimate Std. Error z value Pr(>|z|)  
(Intercept)  2.45984  0.62302  3.948 7.87e-05 ***  
as.factor(DOSE)5 -2.48540  0.64791 -3.836 0.000125 ***  
as.factor(DOSE)20 -1.60201  0.61822 -2.591 0.009561 **  
as.factor(DOSE)80 -1.64440  0.61144 -2.689 0.007158 **  
TIME          -0.04231  0.03267 -1.295 0.195272  
CONC          0.30782  0.10102  3.047 0.002311 **  
---  
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1  
  
Correlation of Fixed Effects:  
              (Intr) a.(DOSE)5 a.(DOSE)2 a.(DOSE)8 TIME  
as.f(DOSE)5 -0.838  
as.(DOSE)20 -0.810  0.713  
as.(DOSE)80 -0.668  0.622   0.643  
TIME         -0.505  0.282   0.239   0.065  
CONC         -0.362  0.222   0.146   -0.185  0.486  
convergence code: 0  
Model failed to converge with max|grad| = 0.00107671 (tol = 0.001, component 1)
```

```
> lrtest(pain.glmm1, pain.glmm)
```

```
Likelihood ratio test
```

```
Model 1: PAINRELIEF ~ as.factor(DOSE) + TIME + CONC + (1 | ID) + CONC  
Model 2: PAINRELIEF ~ as.factor(DOSE) + TIME + CONC + (1 | ID)
```

#DF	LogLik	Df	Chisq	Pr(>Chisq)
1	8	-972.39		
2	7	-1020.66	-1 96.521	< 2.2e-16 ***

```
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

GLMM – Reduced Model

- Removed “TIME” as a covariate (predictor variable)
- All covariates are significant predictors of response
- However, LRT shows reduced model NOT significantly better fit than Full model

```
> lrtest(pain.glmm1, pain.glmm2)
Likelihood ratio test
```

Model 1: PAINRELIEF ~ as.factor(DOSE) + TIME + CONC + (1 | ID) + (1 | CONC)

Model 2: PAINRELIEF ~ as.factor(DOSE) + CONC + (1 | ID) + (1 | CONC)

#Df	LogLik	Df	Chisq	Pr(>Chisq)
1	8	-972.39		
2	7	-973.24	-1	1.6973
				0.1926

```
### Added random effect to CONC and ID, REMOVED 'TIME' --
pain.glmm2<-glmer(PAINRELIEF ~ as.factor(DOSE)+CONC+(1|ID)+(1|CONC),
                     data = pain, family = binomial, REML=F)
summary(pain.glmm2)
```

```
Family: binomial ( logit )
formula: PAINRELIEF ~ as.factor(DOSE) + CONC + (1 | ID) + (1 | CONC)
Data: pain
```

AIC	BIC	logLik	deviance	df.resid
1960.5	1999.4	-973.2	1946.5	1913

caled residuals:

Min	1Q	Median	3Q	Max
2.87699	-0.47419	-0.05387	0.49941	2.90982

andom effects:

Groups	Name	Variance	Std.Dev.
CONC	(Intercept)	0.7614	0.8726
ID	(Intercept)	3.6251	1.9040
umber of obs:	1920, groups:	CONC, 1319; ID, 160	

ixed effects:

	Estimate	Std. Error	z value	Pr(> z)
Intercept)	2.06053	0.53511	3.851	0.000118 ***
s.factor(DOSE)5	-2.25691	0.61826	-3.650	0.000262 ***
s.factor(DOSE)20	-1.41695	0.59717	-2.373	0.017654 *
s.factor(DOSE)80	-1.59958	0.60849	-2.629	0.008569 **
ONC	0.37203	0.08849	4.204	2.62e-05 ***

--
ignif. codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

orrelation of Fixed Effects:

(Intr)	a.(DOSE)5	a.(DOSE)2	a.(DOSE)8
s.f(DOSE)5	-0.840		
s.(DOSE)20	-0.822	0.694	
s.(DOSE)80	-0.738	0.631	0.648
ONC	-0.152	0.099	0.031
			-0.253

Full Model is Final Model

Parameter Estimate Interpretations

1. Exponentiate fixed effects coefficients (slope of predictor variables) to get Odds Ratios

```
exp(fixef(pain.glmm1))
(Intercept) as.factor(DOSE)5 as.factor(DOSE)20 as.factor(DOSE)80      TIME      CONC
11.70295924    0.08329215    0.20149175    0.19312828    0.95856943    1.36045758
```

2. A person given 5mg drug has 0.083x the odds of having pain relief as that same person on placebo

A person given 20mg drug has 0.20x the odds of having pain relief as that same person on placebo

A person given 80mg drug has 0.19x the odds of having pain relief as that same person on placebo

The odds of achieving pain relief increase 1.36x with each unit (1 ng/mL) increase in drug concentration

Comparing GEE and GLMM

- Though not directly comparable, GEE provided odds ratios that made sense
- Odds ratios for predictor variables in GLMM showed a lower odds of having pain relief on each dose of drug vs placebo
- Both models suggested not much additional benefit with 80mg drug vs 20 mg drug
- Account for the random effects (between subject variability)

Random Effects – GLMM Final Model

- Random effects on ID
 - Variance 3.67; standard deviation (square root of variance) of 1.92
 - ***Interpretation:*** subjects at baseline (prior to treatment) have BSV of 1.92 in their probabilities of achieving pain relief
- Random effects on CONC (variability in drug concentration measured between patients at a given time and dose level)
 - Variance 0.81; standard deviation 0.89
 - ***Interpretation:*** there is a standard deviation of 0.89 between subjects in their concentration measurements within a given dose level

Random effects:

Groups	Name	Variance	Std.Dev.
CONC	(Intercept)	0.8082	0.899
ID	(Intercept)	3.6672	1.915

Number of obs: 1920, groups: CONC, 1319; ID, 160

Final Dose Selection

- Apparent dose-proportional increases in drug exposure
- Apparent less-than dose-proportional increases in pain relief
- Marginalized models (using GEE) suggest “average” person would have lower odds of pain relief at 80mg vs 20mg
- GLMM suggests a subject at 80mg would have lower odds of pain relief (vs same subject on placebo) than a subject at 20mg (vs same subject on placebo)
- Final recommendation: 20 mg

Day 4

9:00 - 10:15am:

- Exposure/Response modeling IV
 - Survival data

1:00 –

- Review
- Workshop summary

10:15-10:30am:

- Break

10:30-12pm:

- Cox models

12:00 – 1:00pm:

- Lunch break

Exposure/Response Modeling IV-A

Survival Data Analysis

Cox Proportional Hazards Models

Presented By:

Hechuan Wang, MS

Shamir Kalaria, PharmD

Center for Translational Medicine

UMB

Survival Analysis: Introduction Theory

What is Survival Analysis?

- Class of statistical methods for which the **outcome variable of interest is time** until an event occurs
- Time is measured from the beginning of follow-up until the event occurs or some reason occurs for the observation of time to end (end of trial, dropout, loss to follow up, etc)

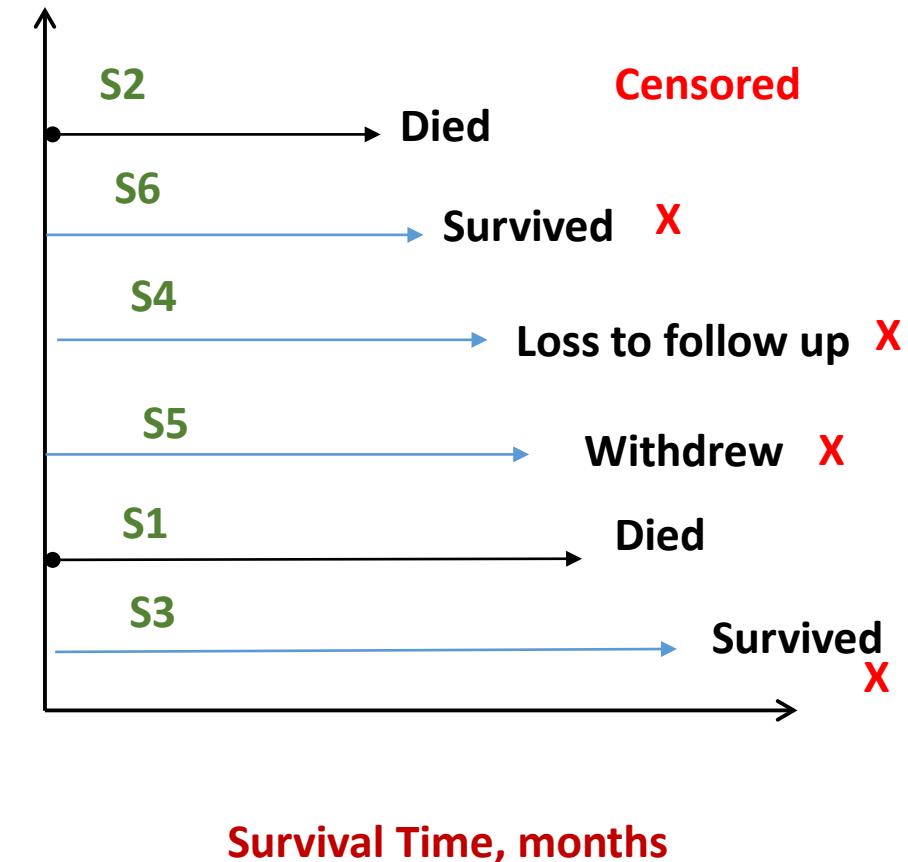
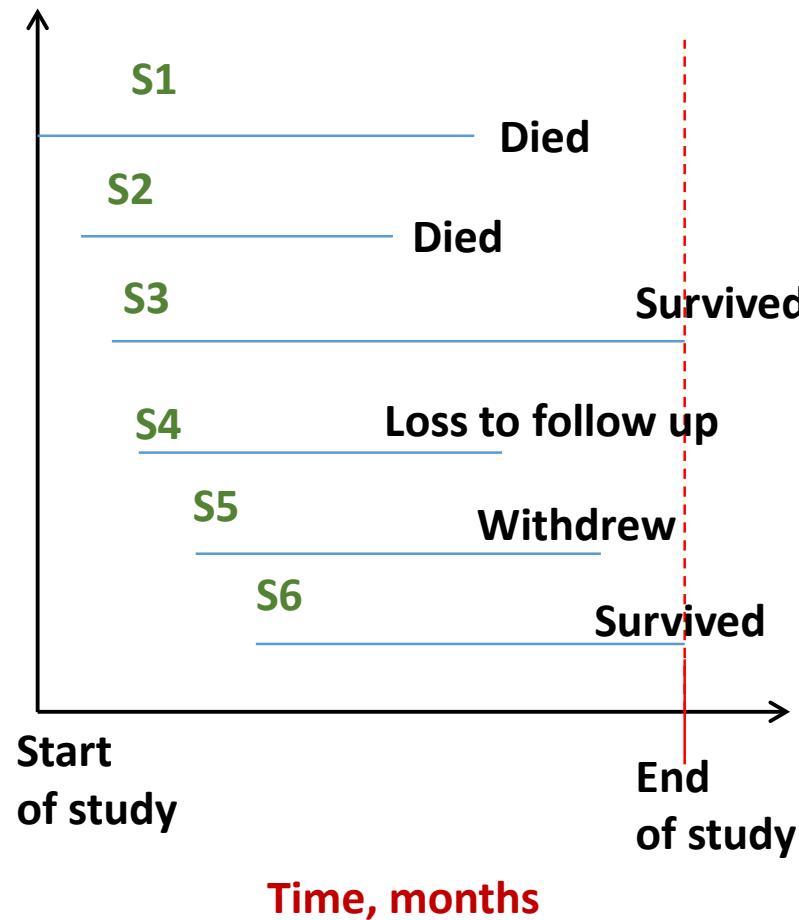
Objective of Survival analysis

- To study **whether an event occurred or not** and the time **when the event occurred**

Examples of survival analysis

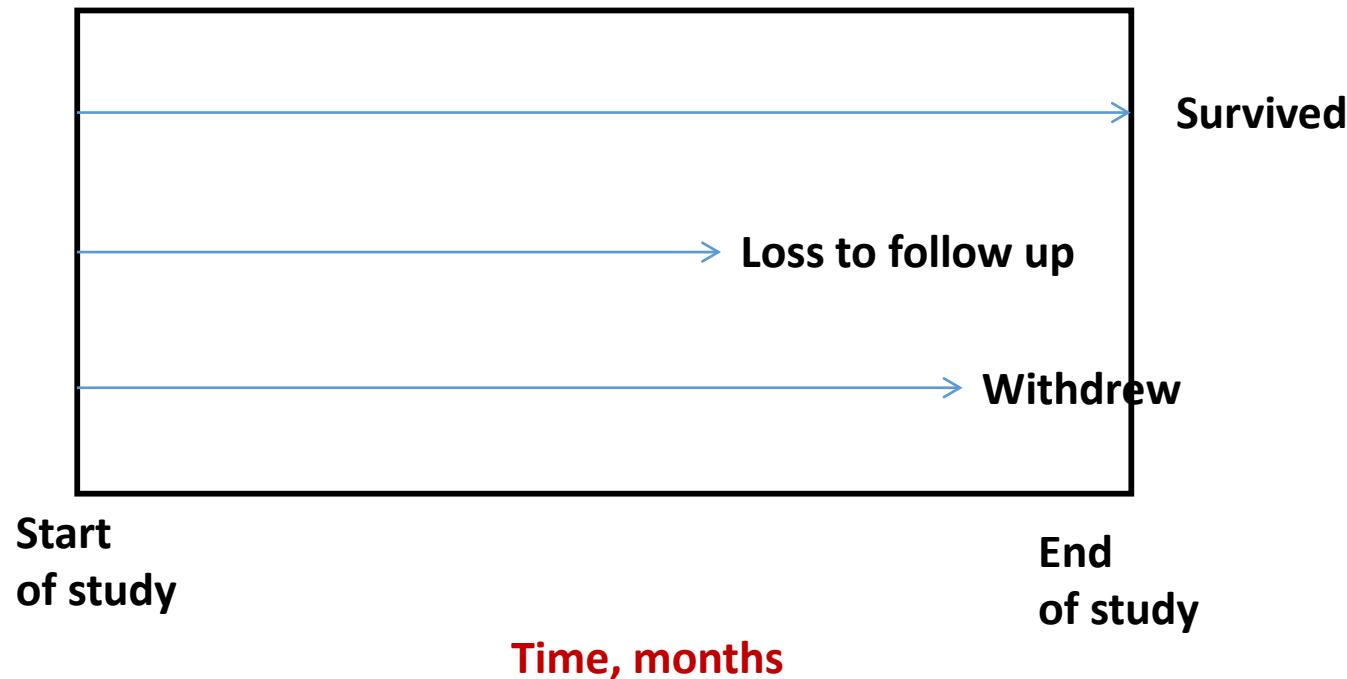
- Cancer trials
 - Follow up of leukemia patients in remission to measure how long they remain in remission
- HIV trials
 - Follow up of patients with HIV infection to measure the time to AIDS
 - Time of AIDS till death

Features of survival data

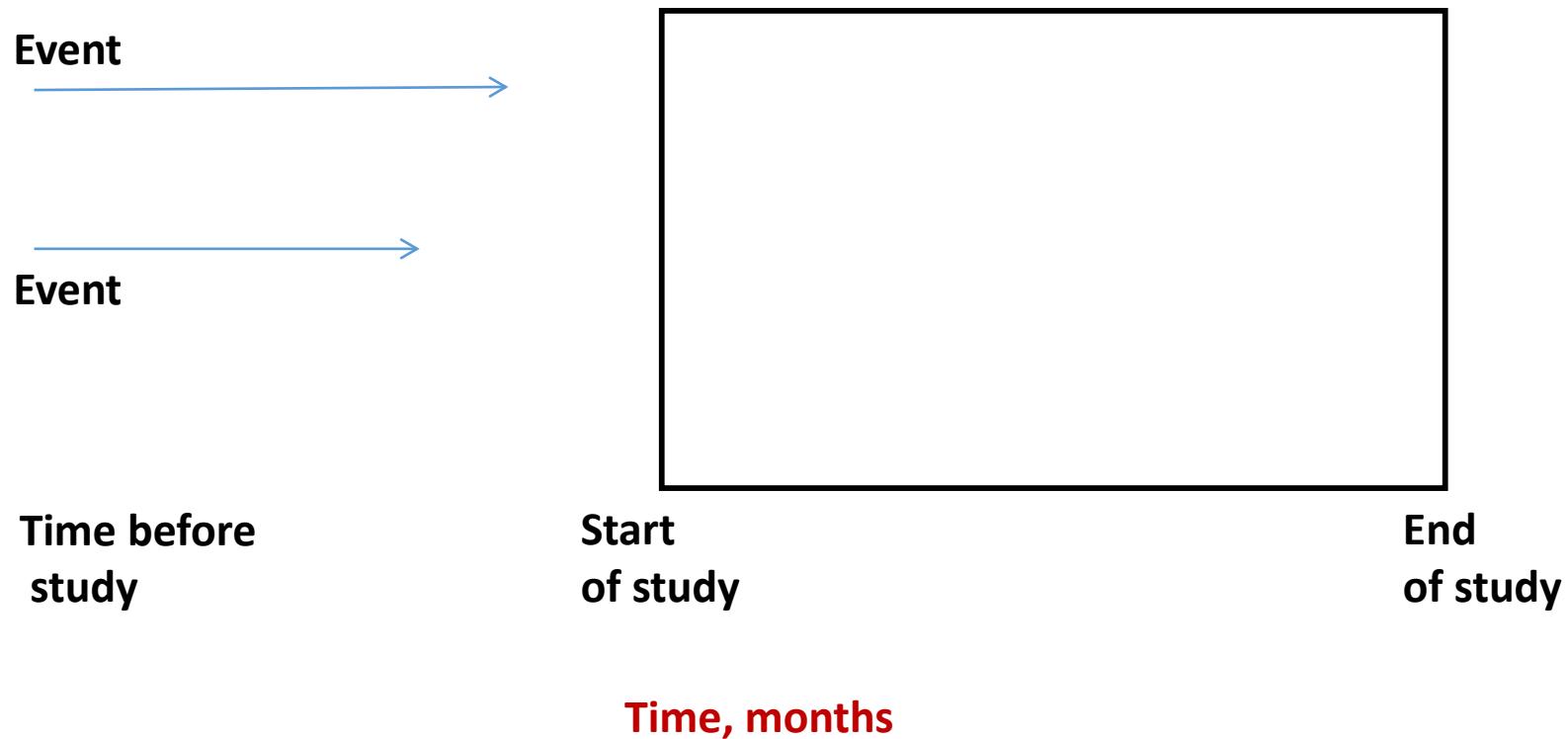


Censoring

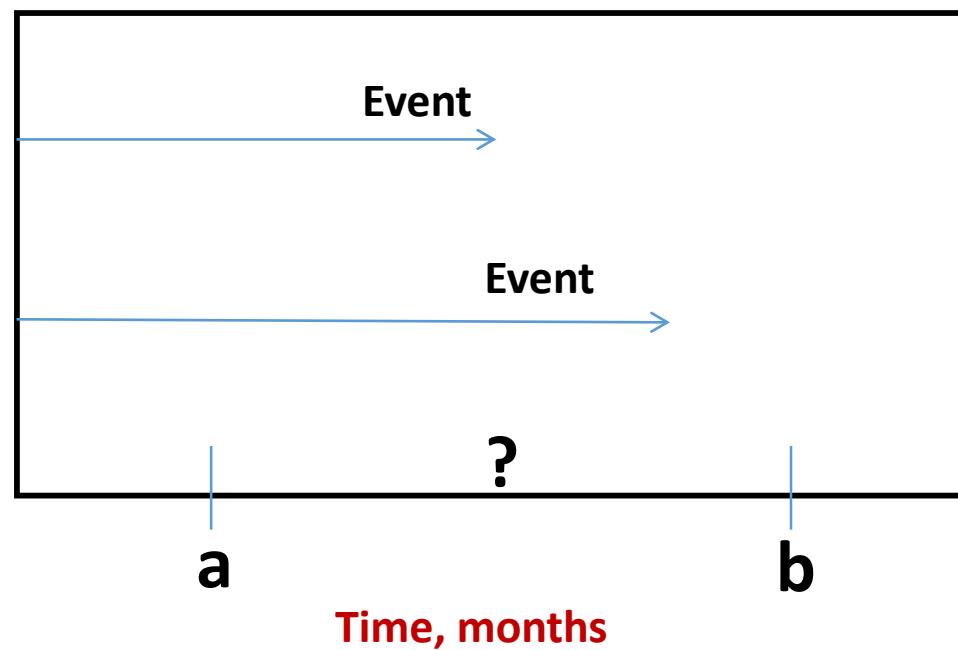
- Three types of censoring
- **Right Censoring**



Left Censoring



Interval Censoring



Main assumption on Censoring

- **Non informative**
 - Subjects who are censored have the same likelihood of experiencing the event had they were not censored
 - Ex: If a patient was lost to follow up, it is assumed that they have the same chance of having the event if they had remained in the trial

Challenges with survival data

- Survival times are always **non-negative**
- Survival times are **censored**
 - Survival time may not be known exactly

Case study – Ovarian cancer dataset

```
> library(survival)  
> data(ovarian)  
> head(ovarian)
```

	futime	fustat	age	resid.ds	rx	ecog.ps
1	59	1	72.3315		2 1	1
2	115	1	74.4932		2 1	1
3	156	1	66.4658		2 1	2
4	421	0	53.3644		2 2	1
5	431	1	50.3397		2 1	1
6	448	0	56.4301		1 1	2

futime: survival or censoring time

fustat: censoring status

age: in years

resid.ds: residual disease present (1=no,2=yes)

rx: treatment group

ecog.ps: ECOG performance status (1 is better)

Summarize dataset

```
> df<-within(ovarian, {  
+   rx<-factor(rx)  
+   ecog.ps<-factor(ecog.ps)  
+   resid.ds<-factor(resid.ds)  
+   fustat<-factor(fustat)  
+ })  
> summary(df )
```

	futime	fustat	age	resid.ds	rx	ecog.ps
Min.	: 59.0	0:14	Min. :38.89	1:11	1:13	1:14
1st Qu.	: 368.0	1:12	1st Qu.:50.17	2:15	2:13	2:12
Median	: 476.0		Median :56.85			
Mean	: 599.5		Mean :56.17			
3rd Qu.	: 794.8		3rd Qu.:62.38			
Max.	:1227.0		Max. :74.50			

Research Questions

- How do the treatments compare with respect to survival benefit?
- How are the predictor variables related to survival times?
 - Does residual disease decrease the survival benefit?
 - Effect of Age on survival risk?

Concepts of survival analysis

- Survival and hazard functions
 - Examine the distribution of survival time
- Assess the relationship of time independent and time dependent covariates to survival time

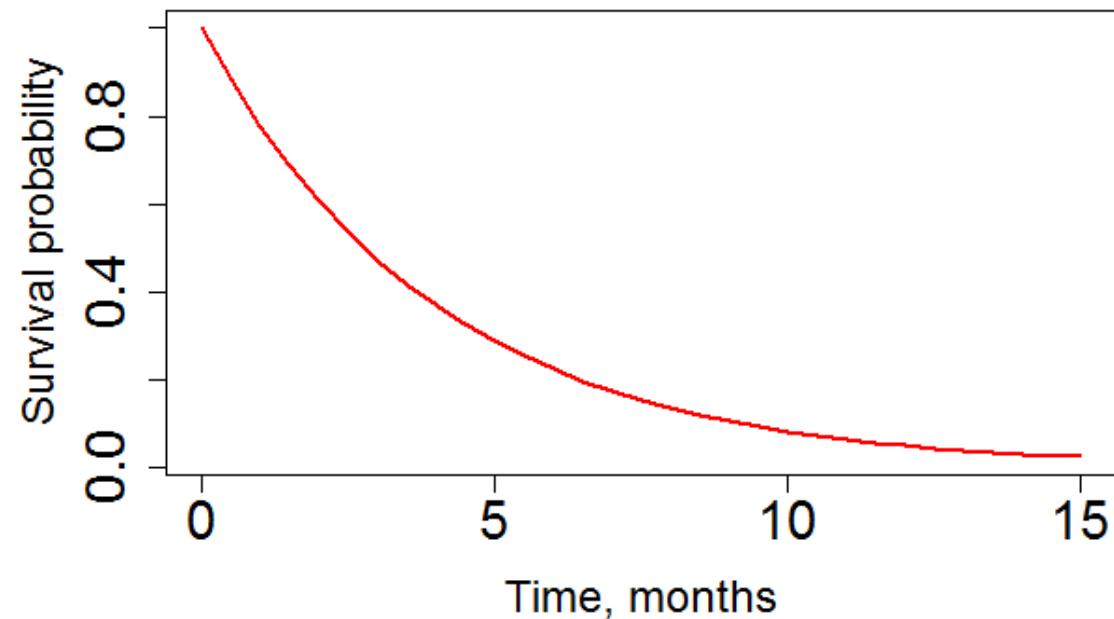
Survival function

- The probability of an individual surviving beyond a time t (experiencing the event after time t)

$$S(t) = P(T > t)$$

$$S(t) = 1 \text{ at } t = 0$$

$$S(t) = 0 \text{ at } t = \infty$$



Hazard function

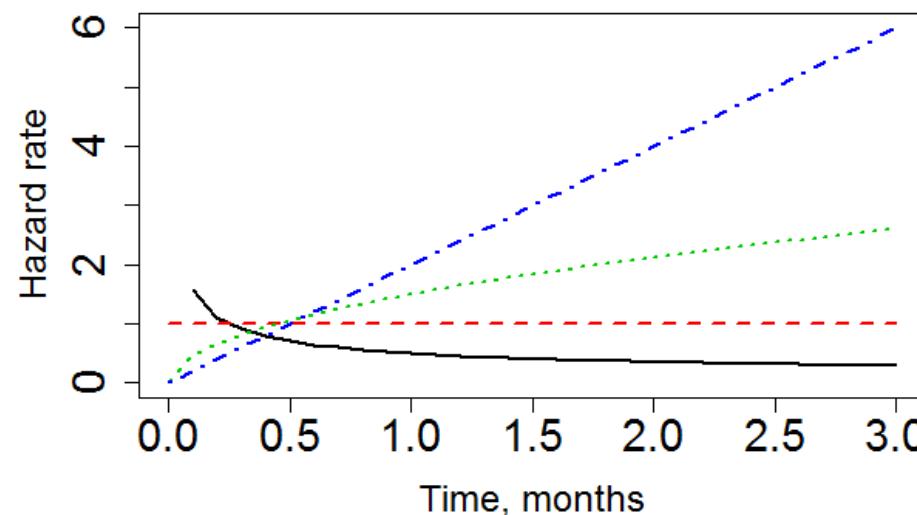
- Instantaneous probability **per unit time** that a patient will experience the event given that they have survived to time t

$$h(t) = \lim_{\Delta t \rightarrow 0} \frac{P[t \leq T \leq t + \Delta t | T \geq t]}{\Delta t}$$

- Numerator of hazard function is a conditional probability
- Hazard function is not a probability, it is a probability rate

Interpretation of hazard function

- Expressed as expected number of events in one-unit interval of time
- Ex: Hazard function is $0.2/\text{day}$ at time t
 - 20% chance to experience the event per day at that moment in time

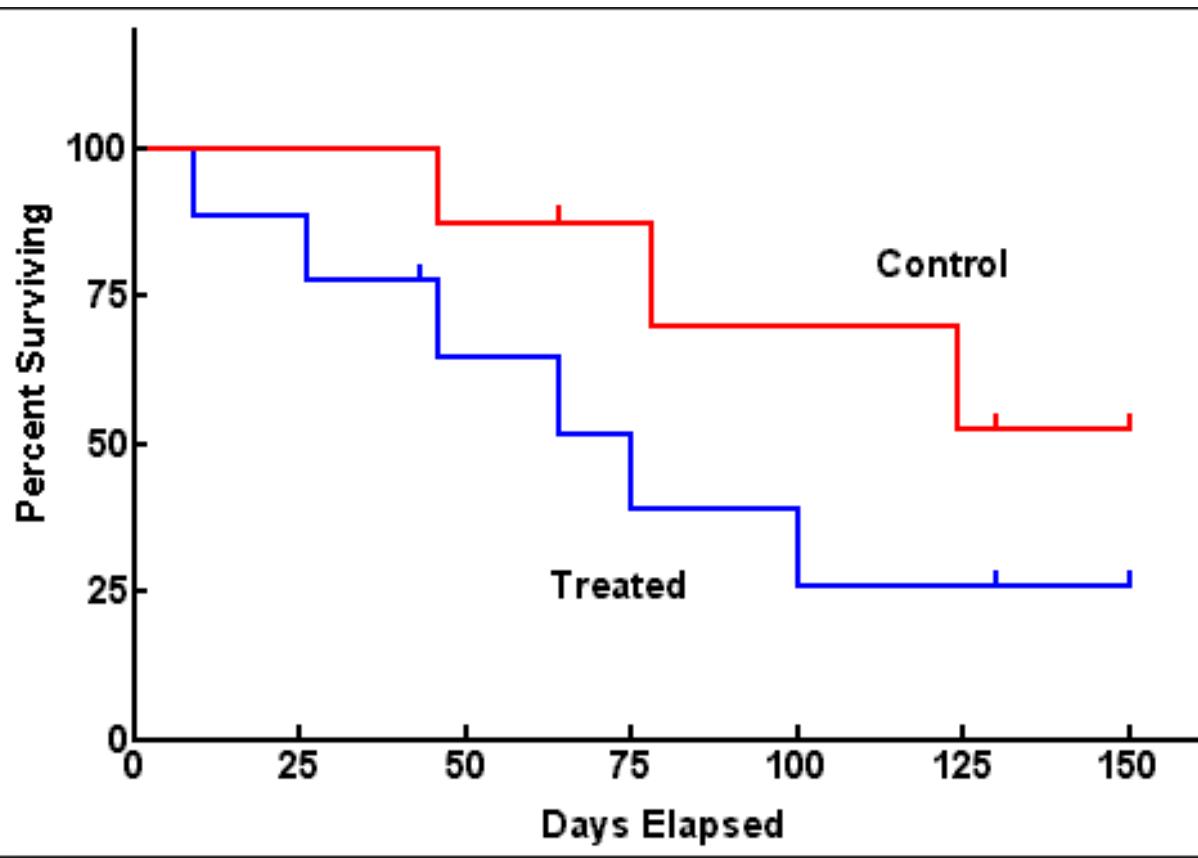


Relationship between hazard function and survival function

- If hazard function is known, then survival function can be derived and vice versa

$$\begin{aligned} S(t) &= \exp\left(-\int_0^t h(u)du\right) \\ &= 1 - F(t) = 1 - P(T \leq t) \\ &= P(T > t) \\ h(t) &= \frac{-dS(t)/dt}{S(t)} = \frac{f(t)}{S(t)} = \frac{P(T = t)}{P(T > t)} \end{aligned}$$

Kaplan-Meier Plots



- Non-parametric method to examine survival probability and to plot the survival function
- Incorporates information from all the observations available, both censored and uncensored to compute the survival probabilities

Kaplan-Meier: Product Limit estimator

- ‘n’ subjects with ordered, independent, distinct survival times

$$t_{(1)}, t_{(2)}, \dots, t_{(n)}$$

$$S(t_j) = P(T > t_j) = S(t_{(j-1)})(1 - \frac{F_j}{n_j})$$



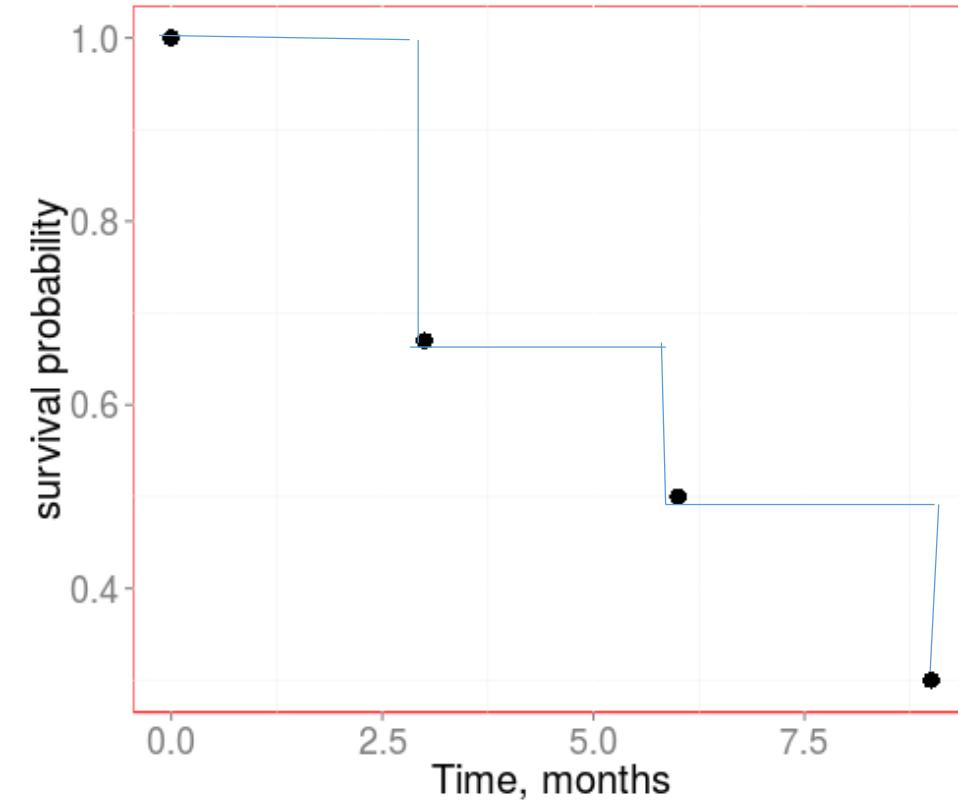
Probability of survival
to time t_j

Probability of
survival
to time t_{j-1}

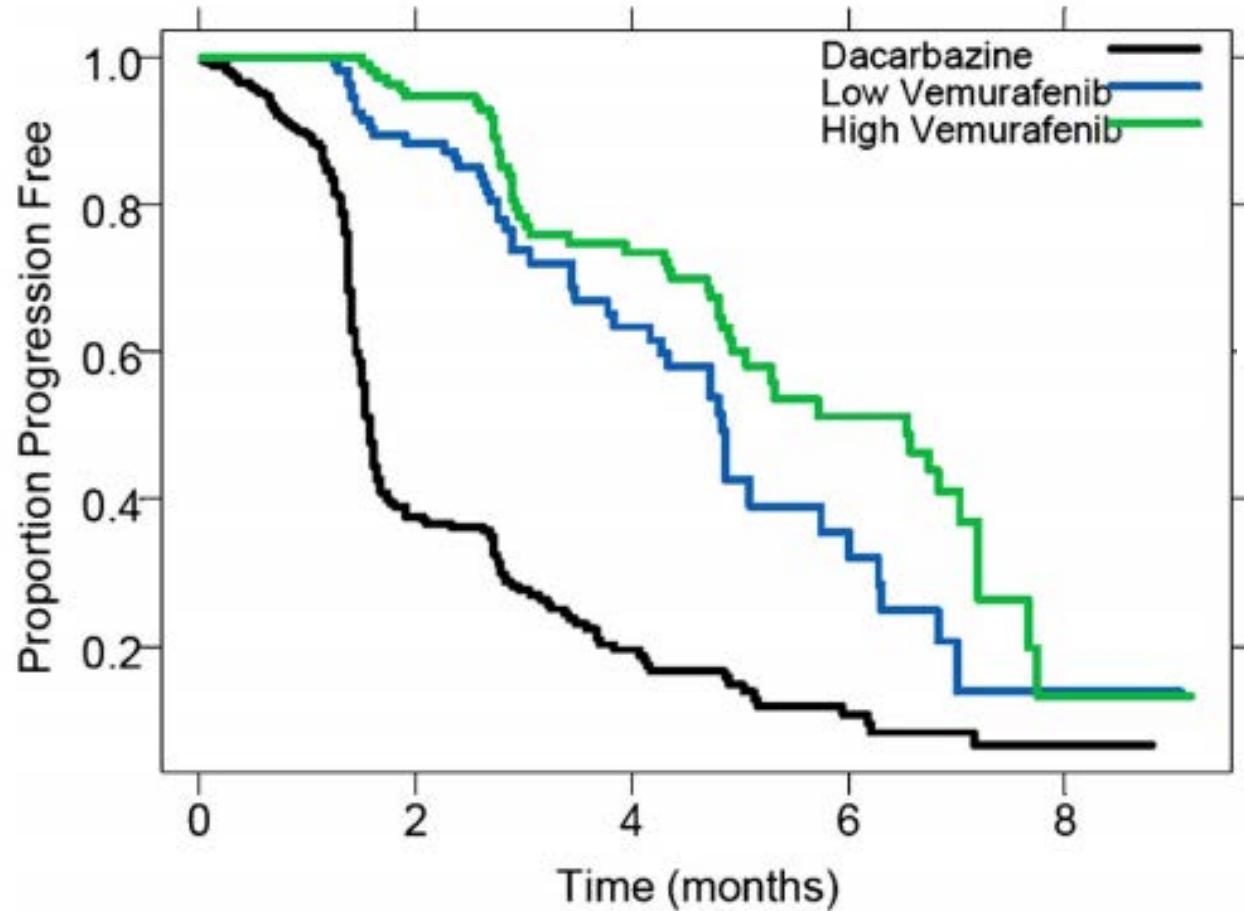
Fraction of subjects
survived till t_j

Kaplan Meier estimator

Time	No at risk	No with events	Survival probability
0	15	0	1
3	10	5	$1 * \left(1 - \frac{5}{15}\right) = 0.67$
6	8	2	$0.67 * \left(1 - \frac{2}{8}\right) = 0.50$
9	5	3	$0.50 * \left(1 - \frac{3}{5}\right) = 0.2$



Compare survival distribution between groups



- Log Rank Test

H_0 : Survival distributions are same among groups

H_a : Survival distributions are different among groups

Test Statistic

$$\frac{\left(\sum_{i=1}^r (d_{1i} - e_{1i})\right)^2}{\text{Var}\left(\sum_{i=1}^r (d_{1i} - e_{1i})\right)} \sim \chi^2_{df=1}$$

Kaplan-Meier plots using R

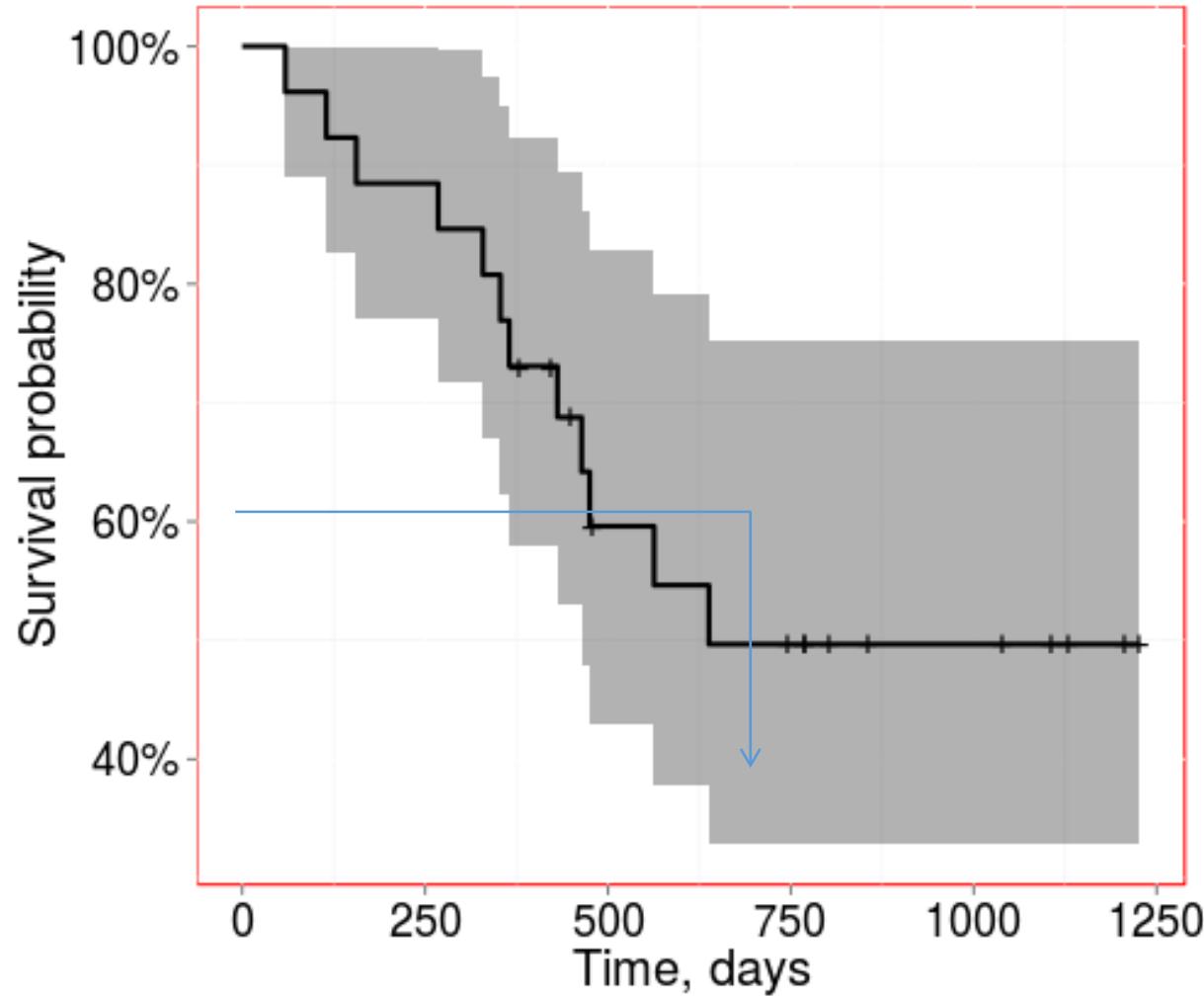
➤ ##### Fit Kaplan-Meier -----

```
> km.fit<-survfit( Surv(futime, fustat)~1, data=ovarian)
> km.fit
Call: survfit(formula = Surv(futime, fustat) ~ 1, data =
ovarian)
```

records	n.max	n.start	events	median	0.95LCL	0.95UCL
26	26	26	12	638	464	NA

- 1: Death or event has occurred
0: Censored

Survival Curve



Summary

```
> summary(km.fit)
Call: survfit(formula = Surv(futime, fustat) ~ 1, data =
ovarian)

  time n.risk n.event survival std.err lower 95% CI upper 95% CI
      59     26      1   0.962  0.0377    0.890  1.000
     115     25      1   0.923  0.0523    0.826  1.000
     156     24      1   0.885  0.0627    0.770  1.000
     268     23      1   0.846  0.0708    0.718  0.997
     329     22      1   0.808  0.0773    0.670  0.974
     353     21      1   0.769  0.0826    0.623  0.949
     365     20      1   0.731  0.0870    0.579  0.923
     431     17      1   0.688  0.0919    0.529  0.894
     464     15      1   0.642  0.0965    0.478  0.862
     475     14      1   0.596  0.0999    0.429  0.828
     563     12      1   0.546  0.1032    0.377  0.791
     638     11      1   0.497  0.1051    0.328  0.752
```

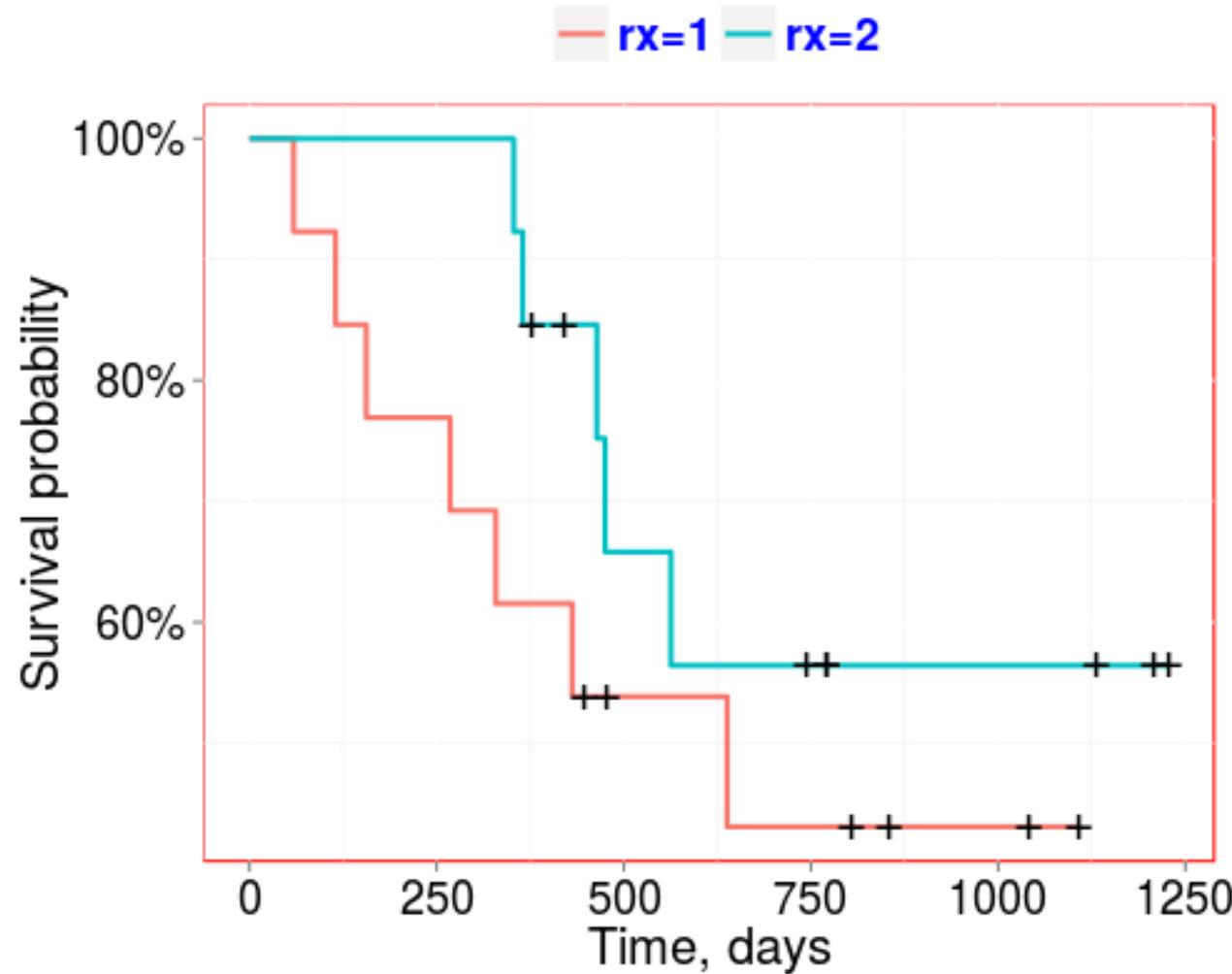
Compare survival between treatments

```
> km.fit_rx<-survfit( Surv(futime, fustat)~rx, data=ovarian)
> km.fit_rx
Call: survfit(formula = Surv(futime, fustat) ~ rx, data = ovarian)
      records n.max n.start events median 0.95LCL 0.95UCL
        rx=1     13     13     13      7     638     268     NA
        rx=2     13     13     13      5      NA     475     NA
```

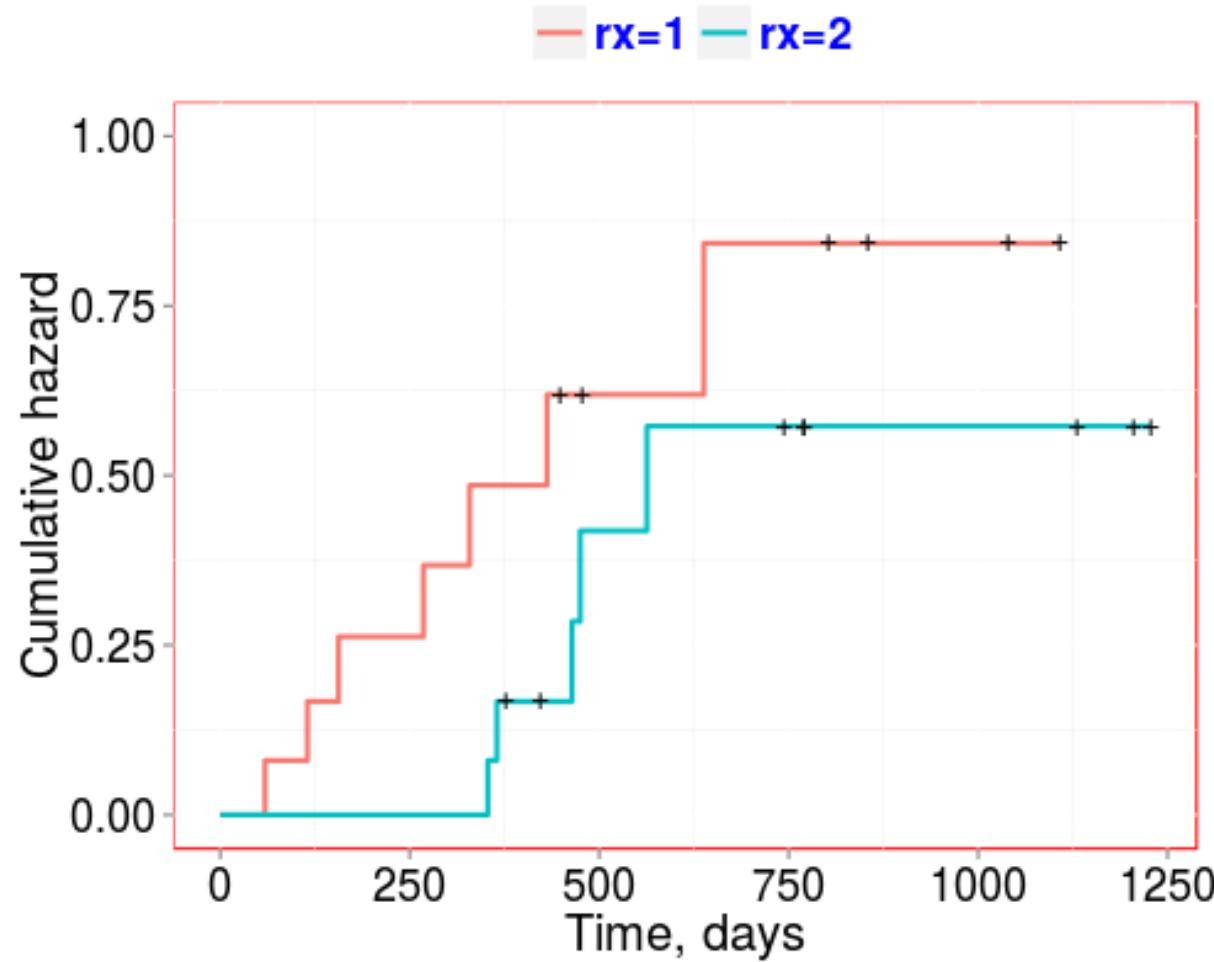
```
> summary(km.fit_rx)
Call: survfit(formula = Surv(futime, fustat) ~ rx, data = ovarian)
```

rx=1									
time	n.risk	n.event	survival	std.err	lower	95% CI	upper	95% CI	
59	13	1	0.923	0.0739		0.789		1.000	
115	12	1	0.846	0.1001		0.671		1.000	
156	11	1	0.769	0.1169		0.571		1.000	
268	10	1	0.692	0.1280		0.482		0.995	
329	9	1	0.615	0.1349		0.400		0.946	

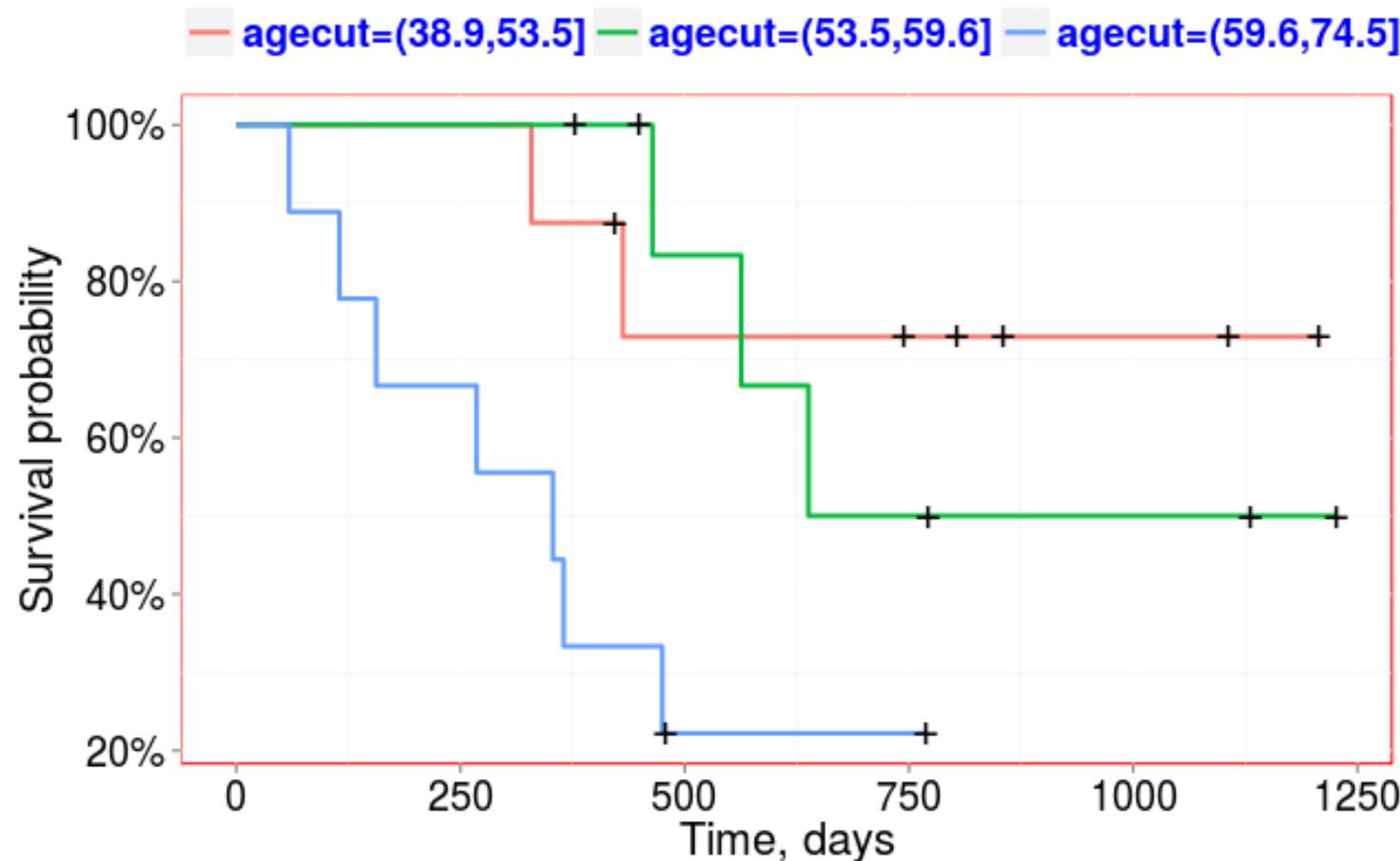
Survival curves for two treatments



Cumulative Hazard function



Survival curves by age categories



Log-Rank test for comparing survival by treatments

H_0 : Survival distributions are same among groups

H_a : Survival distributions are different among groups

```
> test.rx<-survdiff(Surv(futime, fustat)~rx,data = ovarian)
> test.rx
Call:
survdiff(formula = Surv(futime, fustat) ~ rx, data = ovarian)
```

	N	Observed	Expected	$(O-E)^2/E$	$(O-E)^2/V$
rx=1	13	7	5.23	0.596	1.06
rx=2	13	5	6.77	0.461	1.06

Chisq= 1.1 on 1 degrees of freedom, p= 0.303

Log-Rank test for comparing survival by age categories

```
> test.age<-survdiff(Surv(futime, fustat)~agecut,data = ovarian)
> test.age
Call:
survdiff(formula = Surv(futime, fustat) ~ agecut, data = ovarian)
n=25, 1 observation deleted due to missingness.

      N Observed Expected (O-E)^2/E (O-E)^2/V
agecut=(38.9,53.5] 8        2     4.54     1.417    2.302
agecut=(53.5,59.6] 8        3     4.67     0.598    0.987
agecut=(59.6,74.5] 9        7     2.79     6.340    8.558
```

Chisq= 8.7 on 2 degrees of freedom, p= 0.0132

Survival-Exposure relationship

Pediatr Blood Cancer 2004;42:52–58

Methotrexate Pharmacokinetics and Survival in Osteosarcoma[†]

Irene Aquerreta, PharmD, PhD,^{1*} Azucena Aldaz, PharmD, PhD,¹
Joaquín Giráldez, PharmD, PhD,¹ and Luis Sierrasésúmaga, MD, PhD²

The aim of this study was to establish a relationship between the exposure to HDMTX measured either as the AUC or the peak level, and the response of osteosarcoma tumors in terms of patient survival.

Survival-MTX AUC relationship

Pharmacokinetics–Pharmacodynamics Analysis

DFS, and OS were recorded. To compare the influence of AUC in osteosarcoma response, patients were divided in four groups depending on the value of the mean AUC recorded, with 11 patients in each group (group 1: <2,400 µmol/Lhr; group 2: 2,400–3,675 µmol/Lhr; group 3: 3,700–4,800 µmol/L; group 4: >4,800 µmol/Lhr). **Statistical Analysis**

The relationship between exposure to MTX and survival parameters was analyzed by Cox regression. Coxs *F*-test was applied to compare final events of two groups in terms of survival. To compare multiple samples, a non-parametric test that is an extension of Gehans generalized Wilcoxon test, Peto and Petos generalized Wilcoxon test, and the log-rank test was used. To compare multiple samples two by two, Coxs *F*-test with Bonferroni adjustment [28] was used. Cumulative proportional survival graphs were realized by Kaplan–Meier curves. The statis-

Significant Disease free survival – AUC relationship

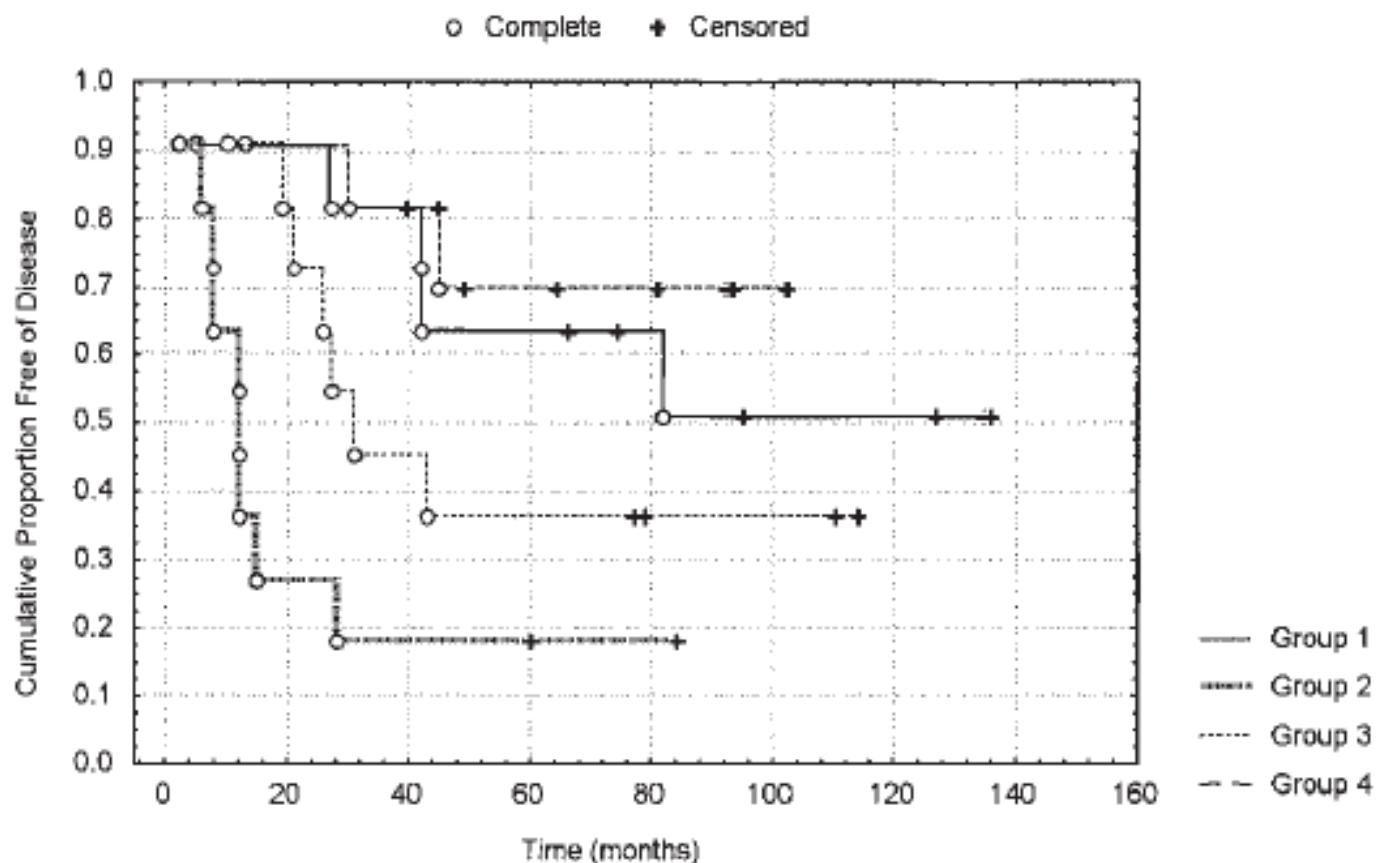


Fig. 1. Disease-free survival (DFS) in the mean area under the concentration–time curve (AUC) groups. Group 1: mean AUC < 2,400 $\mu\text{mol}/\text{Lhr}$; group 2: mean AUC 2,400–3,675 $\mu\text{mol}/\text{Lhr}$; group 3: mean AUC 3,700–4,800 $\mu\text{mol}/\text{L}$; group 4: mean AUC > 4,800 $\mu\text{mol}/\text{Lhr}$.

Overall survival – MTX AUC relationship

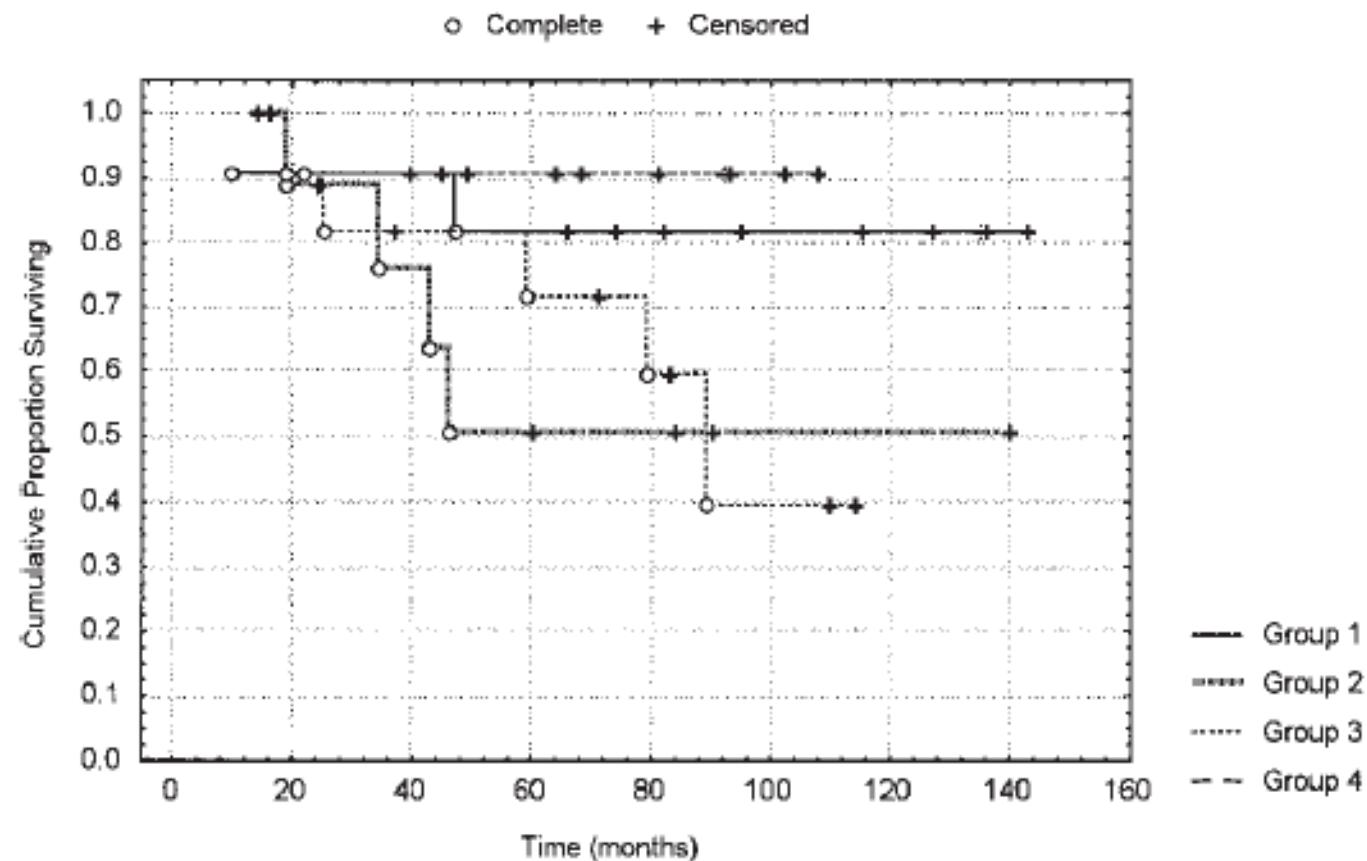


Fig. 2. Overall survival (OS) in the mean AUC groups. Group 1: mean AUC < 2,400 $\mu\text{mol/Lhr}$; group 2: mean AUC 2,400–3,675 $\mu\text{mol/Lhr}$; group 3: mean AUC 3,700–4,800 $\mu\text{mol/L}$; group 4: mean AUC > 4,800 $\mu\text{mol/Lhr}$.

Survival models

	Parametric	Non-parametric
Distribution of survival time	Assumed to be known	Unknown
Hazard function	Completely specified	Unspecified
Examples	Exponential, Weibull, log-normal models	Cox-proportional hazards model

$$\begin{aligned}S(t) &= \exp\left(-\int_0^t h(u)du\right) \\&= 1 - F(t) = 1 - P(T \leq t) \\&= P(T > t)\end{aligned}$$

$$h(t) = \frac{-dS(t)/dt}{S(t)} = \frac{f(t)}{S(t)} = \frac{P(T = t)}{P(T > t)}$$

Cox Model, 1972-a quick history

Regression Models and Life-Tables

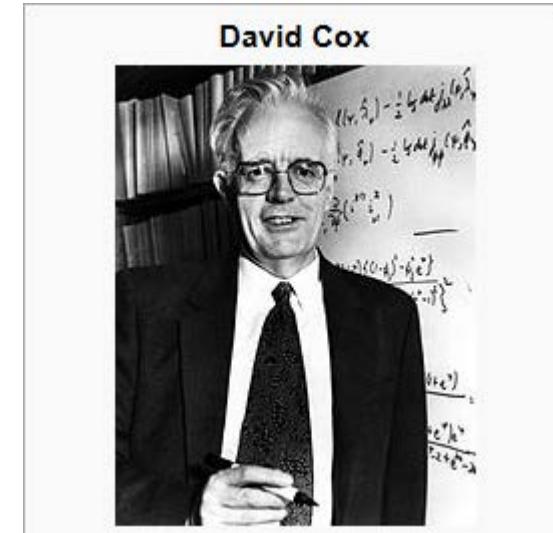
By D. R. Cox

Imperial College, London

[Read before the ROYAL STATISTICAL SOCIETY, at a meeting organized by the Research Section, on Wednesday, March 8th, 1972, Mr M. J. R. HEALY in the Chair]

SUMMARY

The analysis of censored failure times is considered. It is assumed that on each individual are available values of one or more explanatory variables. The hazard function (age-specific failure rate) is taken to be a function of the explanatory variables and unknown regression coefficients multiplied by an arbitrary and unknown function of time. A conditional likelihood is obtained, leading to inferences about the unknown regression coefficients. Some generalizations are outlined.



David Cox

- the most cited paper in the whole history of JRSS
- the third most cited paper in medical journals
- it has a total of nearly **30,000 citations** (according to Web of Science)
- and this is still increasing

Cox proportional hazards (CPH) model

- Hazard of the event at time t is a product of baseline hazard function $h_o(t)$ and the exponent of linear sum of the predictor variables

$$h(t; x) = h_o(t) \exp(\beta_1 x_1 + \beta_2 x_2 + \dots + \beta_k x_k)$$

Baseline hazard function
Only a function of time

Linear function of the predictors
Time-independent covariates

$$\log(h(t; x)) = \log(h_o(t)) + \beta_1 x_1 + \beta_2 x_2 + \dots + \beta_k x_k$$

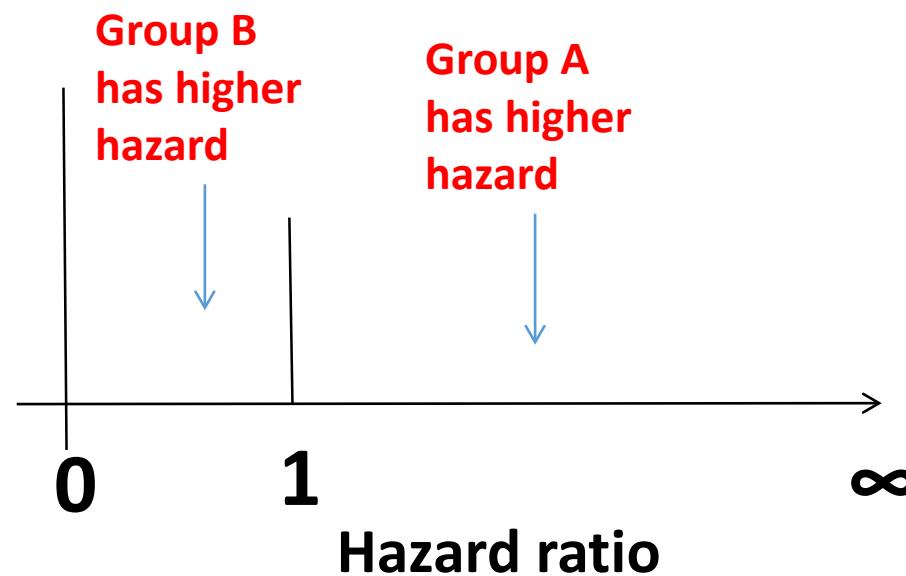
Hazard Ratio

- Key concept in Cox regression
- Hazards are proportional over time : Hazard ratio is constant over time

$$\begin{aligned}\text{HazardRatio} &= \frac{\text{Hazard in group A}}{\text{Hazard in group B}} \\ &= \frac{h_o(t) \exp(\beta_i x_{iA})}{h_o(t) \exp(\beta_i x_{iB})} \\ &= \exp(\beta_i(x_{iA} - x_{iB}))\end{aligned}$$

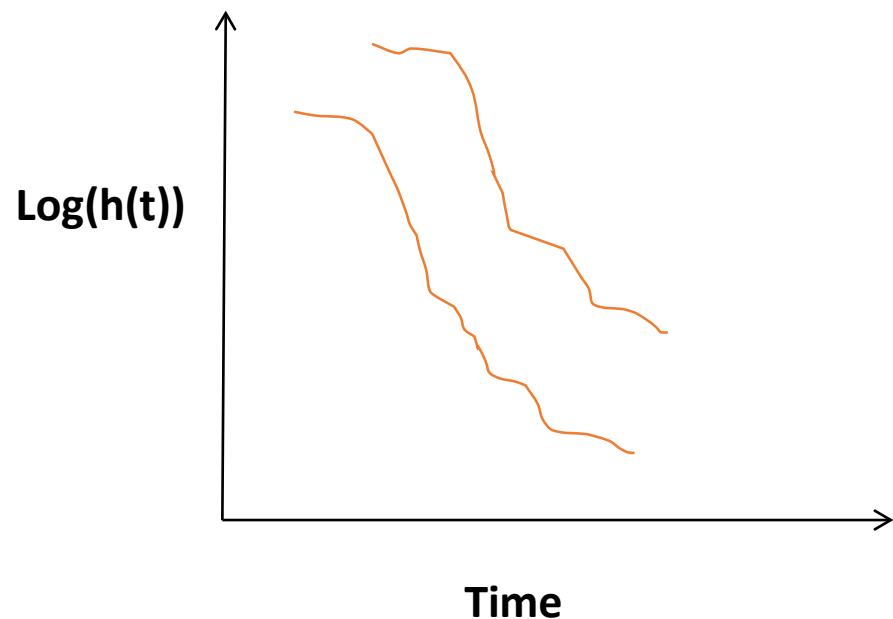
Properties of Hazard Ratio

- **Positive number:** ranges from 0 to infinity
- **HR of 1:** **No association** between the hazard of the event and the predictor variable



Proportional odds assumption

- Hazards for one group is proportional to the hazard of another group
- Estimated Hazard ratio is constant over time



Estimation of parameters of CPH model

- By Partial likelihood methods
 - Only considers probabilities of subjects who had an event
 - Considers censored observations as part of the risk set until they are censored
 - Baseline hazard function need not be specified
- Inference on maximum partial likelihood parameters: LRT's or T-tests can be used

Illustration of partial likelihood

Subject	Survival time	Censor status
A	2.0	1
B	3.0	0
C	4.0	1
D	5.0	0

$$L_A = \frac{h_A(2)}{h_A(2) + h_B(2) + h_C(2) + h_D(2)}$$

$$L_C = \frac{h_C(4)}{h_C(4) + h_D(4)}$$

Why baseline hazard function need not be specified?

- Baseline hazard function is common in the partial likelihood and hence cancels out

$$L_C = \frac{h_0(4) \exp(\beta_1 C_1 + \beta_2 C_2 + \dots + \beta_k C_k)}{h_0(4) \exp(\beta_1 C_1 + \beta_2 C_2 + \dots + \beta_k C_k) + h_0(4) \exp(\beta_1 D_1 + \beta_2 D_2 + \dots + \beta_k D_k)}$$

$$L(\beta) = L_A \times L_C$$

AM Break

Survival Analysis

Cox-proportional hazards model: Application

Case study – Ovarian dataset

```
##### Cox proportional hazards model -----
cph.fit <- coxph(Surv(futime, fustat) ~ rx+resid.ds+
ecog.ps+age, data = ovarian, ties="breslow")
> summary(cph.fit)

Call:
coxph(formula = Surv(futime, fustat) ~ rx + resid.ds + ecog.ps
+age, data = ovarian, ties = "breslow")

n= 26, number of events= 12

            coef  exp(coef)  se(coef)      z  Pr(>|z| )
rx       -0.91450   0.40072   0.65332 -1.400  0.16158
resid.ds  0.82619   2.28459   0.78961  1.046  0.29541
ecog.ps   0.33621   1.39964   0.64392  0.522  0.60158
age       0.12481   1.13294   0.04689  2.662  0.00777 **

---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

CPH model in R

$H_o : \beta_1 = \beta_2 = \dots = \beta_k = 0$ vs

H_a : At least one of the coefficients is not zero

	exp(coef)	exp(-coef)	lower .95	upper .95
rx	0.4007	2.4955	0.1114	1.442
resid.ds	2.2846	0.4377	0.4861	10.738
ecog.ps	1.3996	0.7145	0.3962	4.945
age	1.1329	0.8827	1.0335	1.242

Concordance= 0.807 (se = 0.091)

Rsquare= 0.481 (max possible= 0.932)

Likelihood ratio test= 17.04 on 4 df, p=0.001896

Wald test = 14.25 on 4 df, p=0.006538

Score (logrank) test = 20.81 on 4 df, p=0.0003449

Significance of Age

```
> #####Test the significance of age-----
> cph.fit.noage <- coxph(Surv(futime, fustat) ~ rx+resid.ds+ecog.ps,data =
ovarian, ties="breslow" )
> cph.fit.noage$loglik
[1] -34.98494 -31.96961
> cph.fit$loglik
[1] -34.98494 -26.46329

> ## Partial loglikelihood ratio test
> lrt.age = 2*(cph.fit$loglik[2] - cph.fit.noage$loglik[2])

lrt.age
[1] 11.01264      Compare with Chi-squared critical value of 3.84 at df=1
```

Final Hazard Model

- Hazard of survival from ovarian cancer as a function of age

$$h(t) = h_o(t) \exp(\beta \times \text{Age})$$

```
> ##### Final model -----
> cph.fit.age<-coxph(Surv(futime, fustat) ~ age,
+                         data = ovarian,method="breslow" )
> summary(cph.fit.age)

      coef  exp(coef)    se(coef)      z Pr(>|z|)    
age 0.16162  1.17541  0.04974  3.249  0.00116 ** 
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

	exp(coef)	exp(-coef)	lower .95	upper .95
age	1.175	0.8508	1.066	1.296

Hazard Ratio interpretation: Age as a continuous variable

- Hazard ratio for age (continuous variable) is 1.17
 - For a one unit increase in age, the hazard of survival from ovarian cancer (death) increases by 17%

Age as a categorical predictor

```
> ##### Hazard Ratio by Age categories-----
> ##### Stratify by age-----
> ovarian$agecateg<-ifelse(ovarian$age <60, "Age <60", "Age > 60")
>
> cph.fit.agecateg<-coxph(Surv(futime, fustat) ~ agecateg,
+                         data = ovarian,method="breslow")
> summary(cph.fit.agecateg)
Call:
coxph(formula = Surv(futime, fustat) ~ agecateg, data = ovarian,
      method = "breslow")

n= 26, number of events= 12



|          | coef     | exp(coef) | se(coef) | z      | Pr(> z )          |
|----------|----------|-----------|----------|--------|-------------------|
| agecateg | Age > 60 | 2.3260    | 10.2369  | 0.6732 | 3.455 0.00055 *** |
|          | ---      |           |          |        |                   |


Signif. codes: 0 `***' 0.001 `**' 0.01 `*' 0.05 `.' 0.1 ` ' 1



|          | exp(coef) | exp(-coef) | lower .95 | upper .95  |
|----------|-----------|------------|-----------|------------|
| agecateg | Age > 60  | 10.24      | 0.09769   | 2.736 38.3 |



Concordance= 0.743 (se = 0.057 )
Rsquare= 0.359 (max possible= 0.932 )
Likelihood ratio test= 11.55 on 1 df, p=0.0006766
Wald test = 11.94 on 1 df, p=0.0005501
Score (logrank) test = 17.09 on 1 df, p=3.561e-05
```

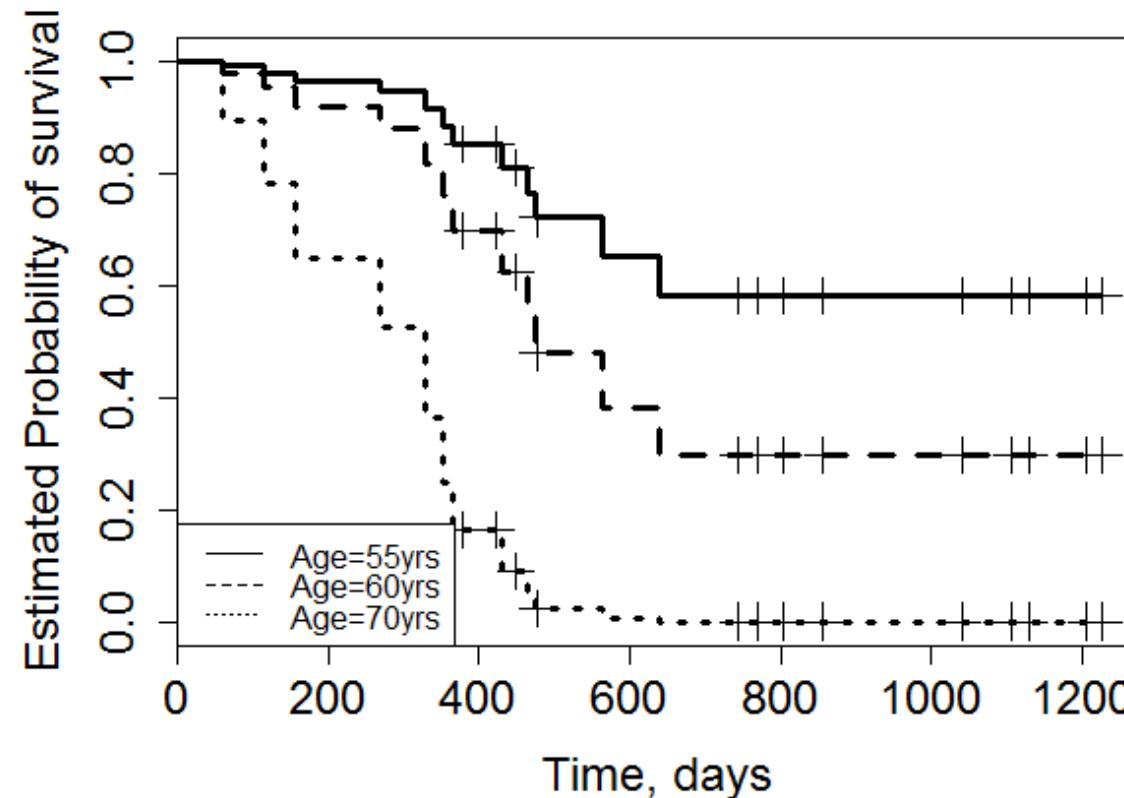
Hazard Ratio interpretation:

Age as a categorical predictor

- Hazard ratio for Age > 60 = 10.2
- The hazard of survival from ovarian cancer (death) is about 10 times likely in patients aged 60 and above as compared to subjects less than 60 years of age.

Estimated survival function

- Based on the cox model, survival function can be estimated for a specific covariate value



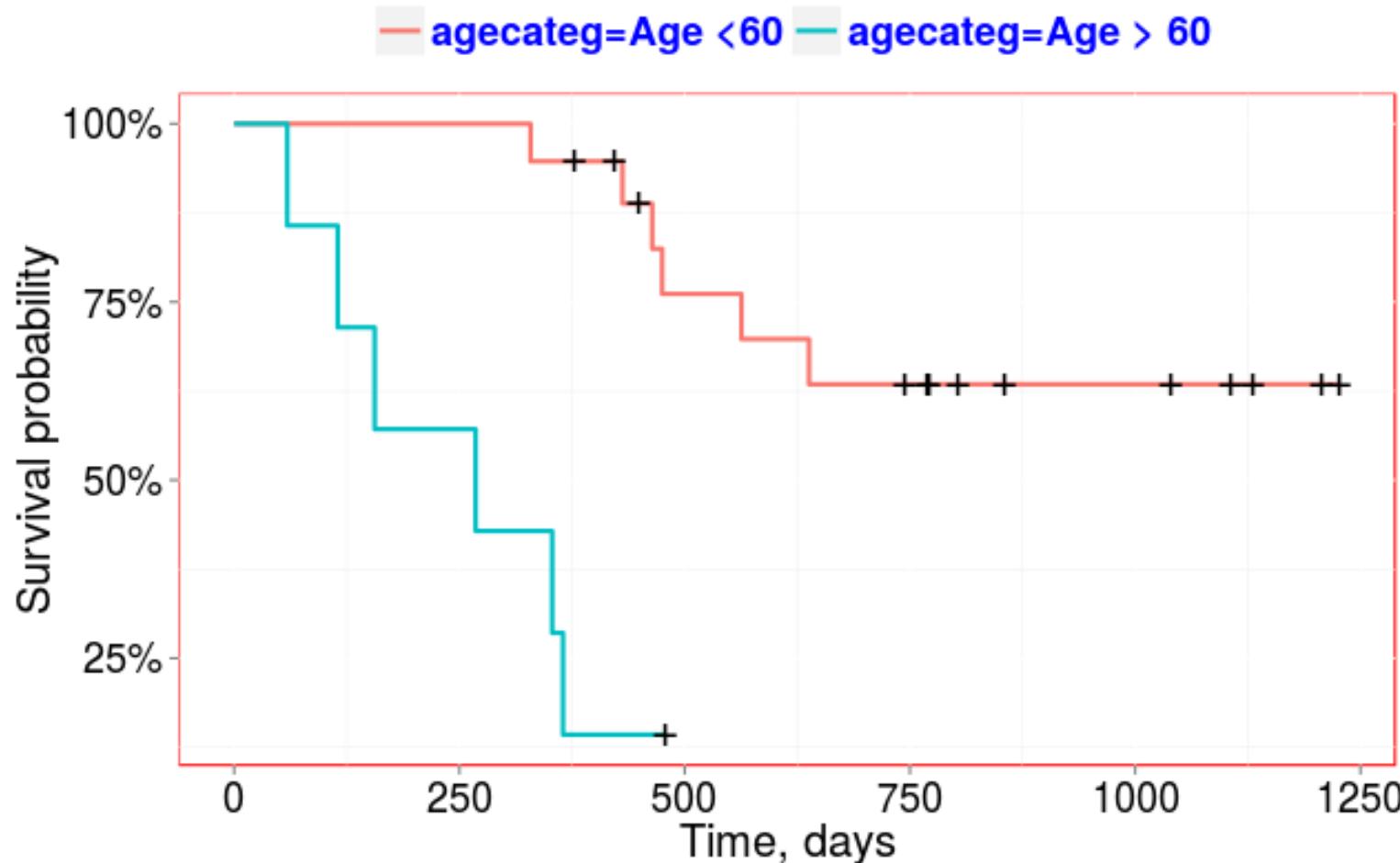
Model Assessment

- Check the proportional hazards assumption
 - Parallel Kaplan Meier plots
 - Schoenfeld residuals
 - Global Chi-squared test
- Overall goodness of fit
 - Cox-Snell residuals
- Residual plots
 - Outlier and influential points detection
 - Deviance residuals, DFBETA

Check Proportional hazards assumption

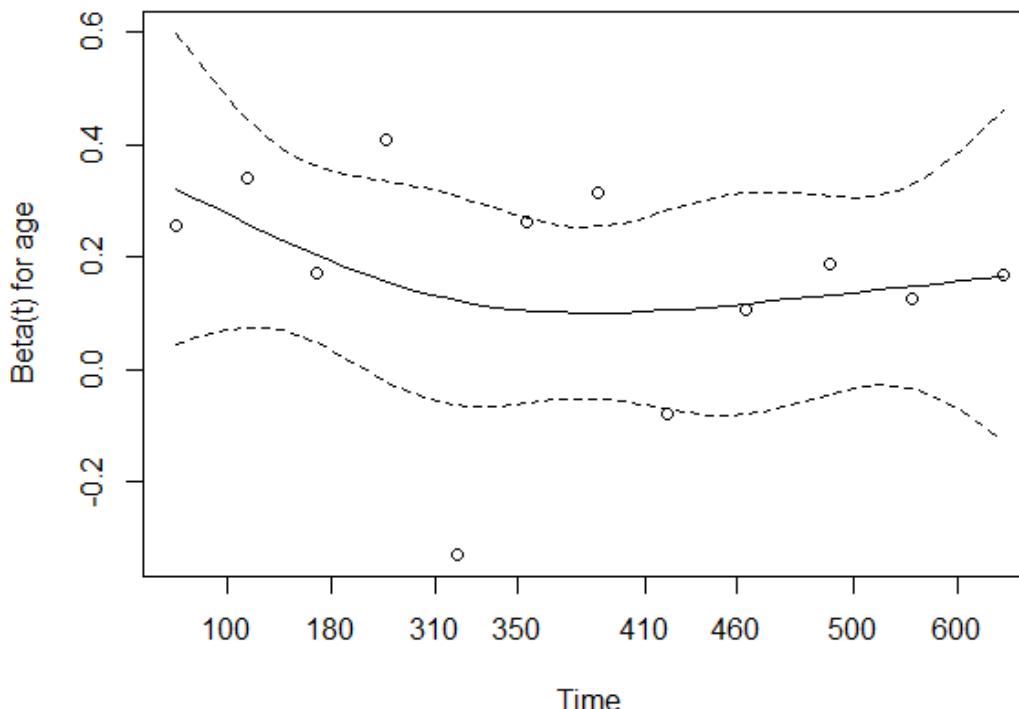
- Graphical method
 - Kaplan Meier plots should look **parallel** with respect to groups of covariates
 - Categorical covariates
 - Continuous covariates : stratify them
- Global Chi-squared test

PH assumption for Age: Ovarian dataset



Global Chi-squared test

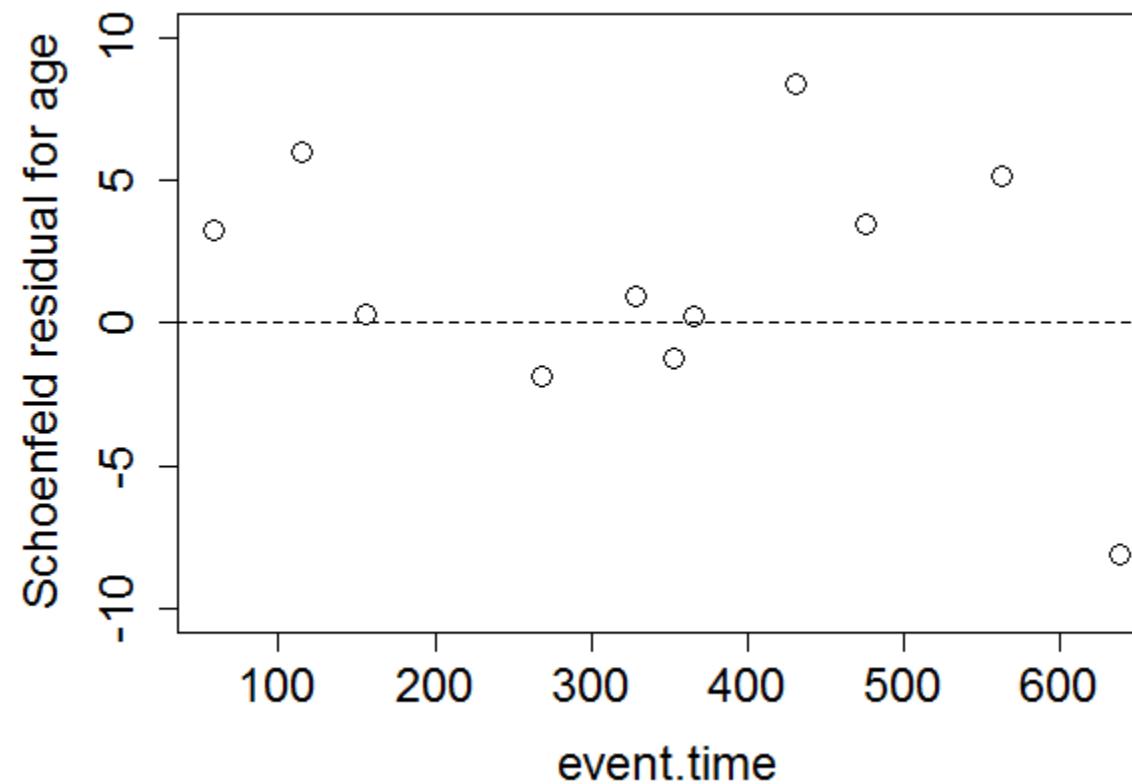
```
> ##### Cox proportional hazards assumption-----
> temp <- cox.zph(cph.fit.age)
> temp
      rho  chisq      p
age -0.209  0.647  0.421
> plot(temp)
```



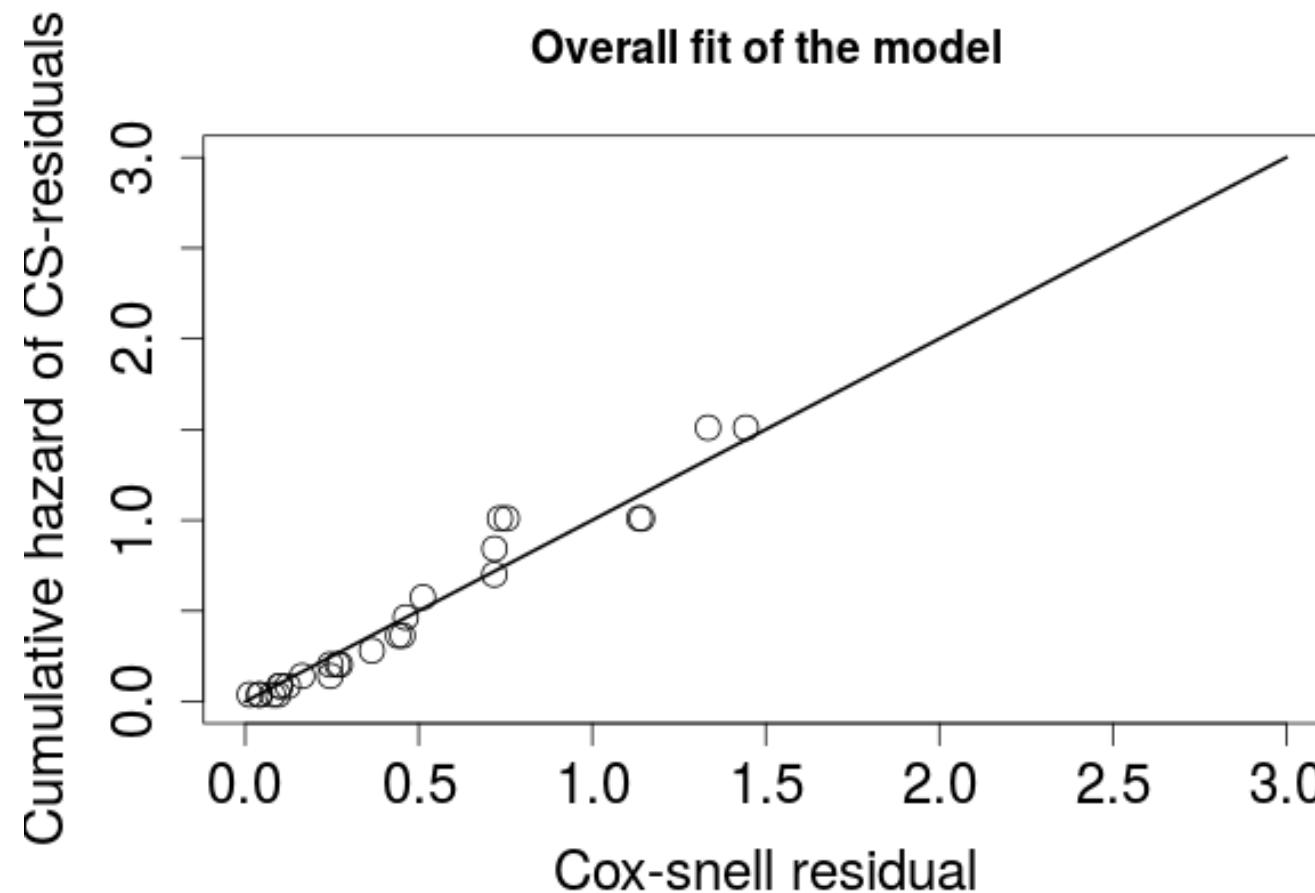
Output Residuals for model assessment

```
> ##### Types of Residuals-----
> mtg.1 = resid(cph.fit.age,type="martingale" )
> coxsn.1 = ovarian$fustat - mtg.1
> dev.1 = resid(cph.fit.age,type="deviance" )
> schoen.1 = resid(cph.fit.age,type="schoenfeld" )
```

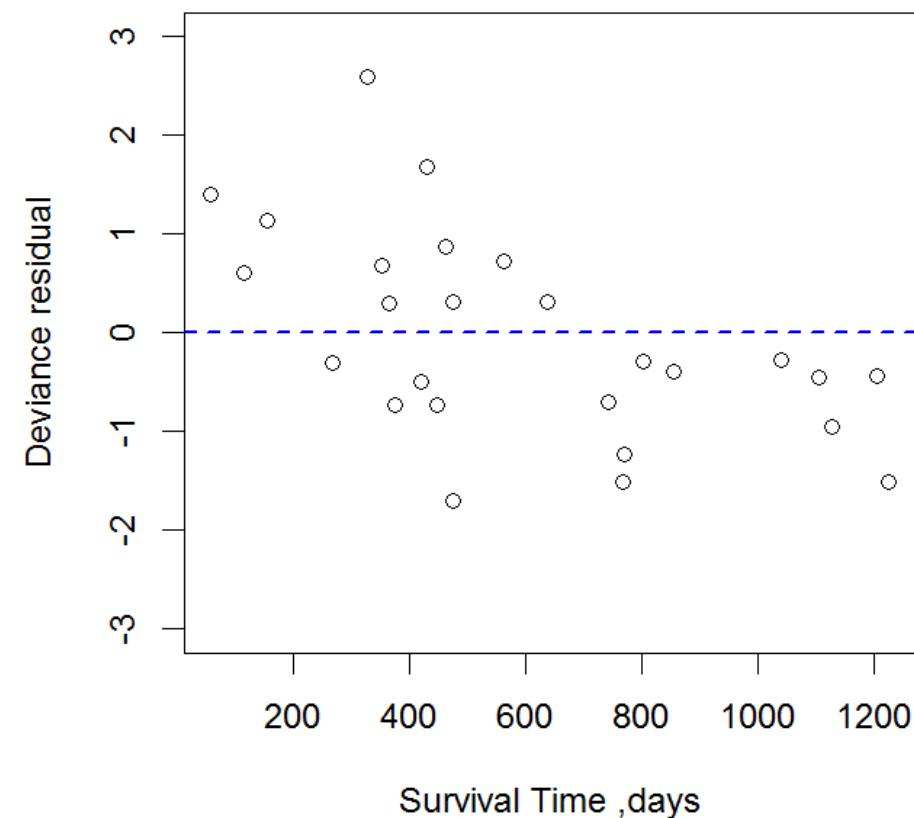
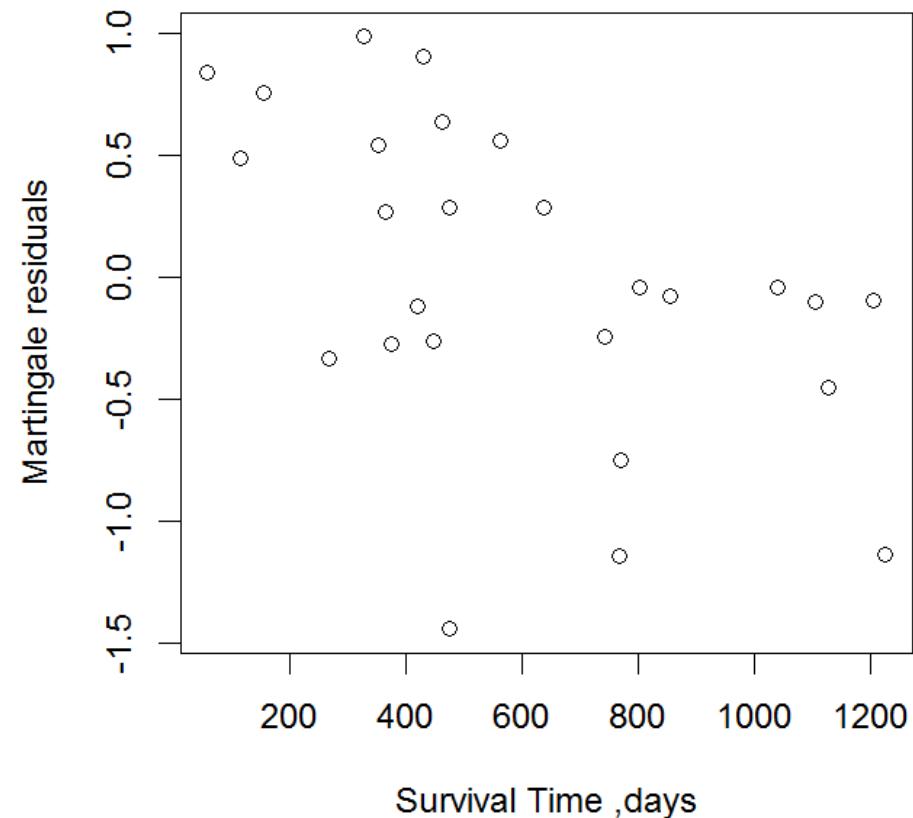
Schoenfeld residuals



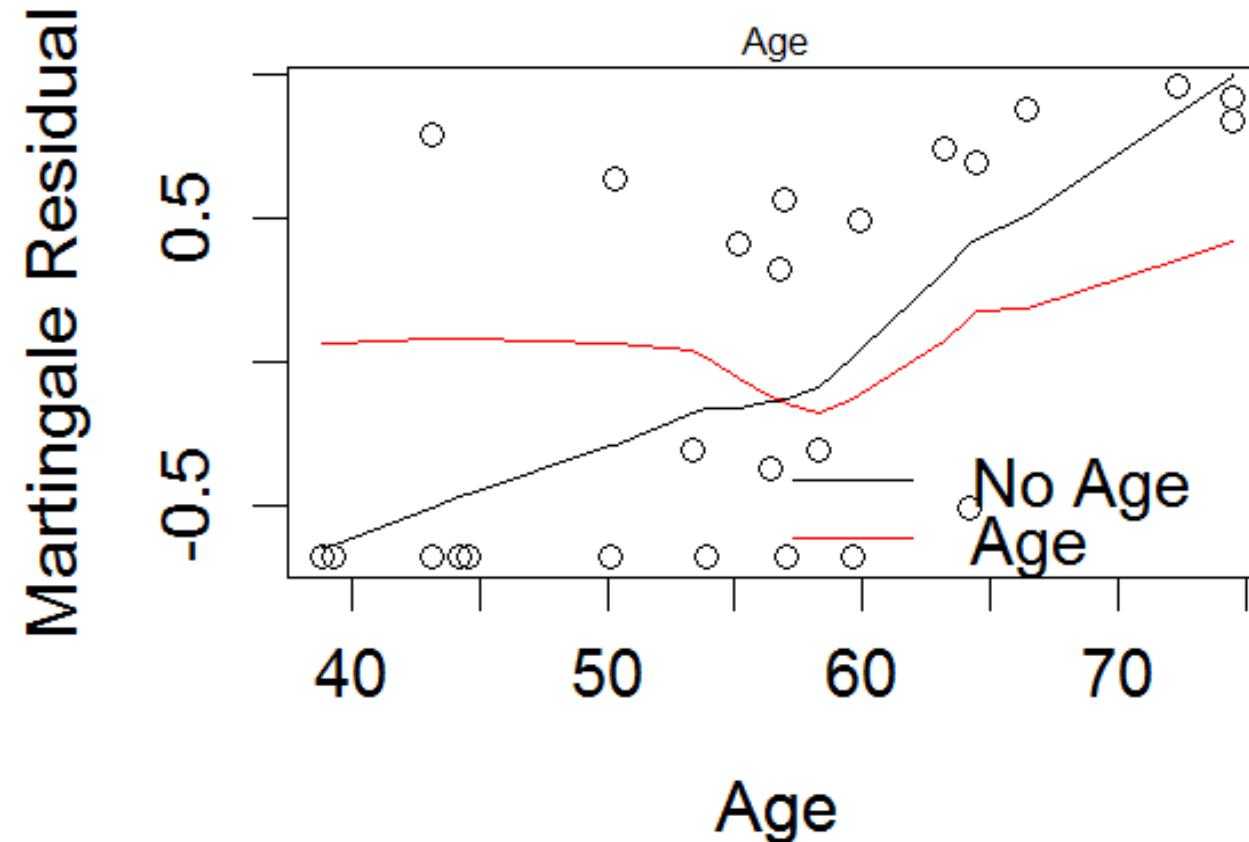
Overall model fit: Cox-Snell residuals



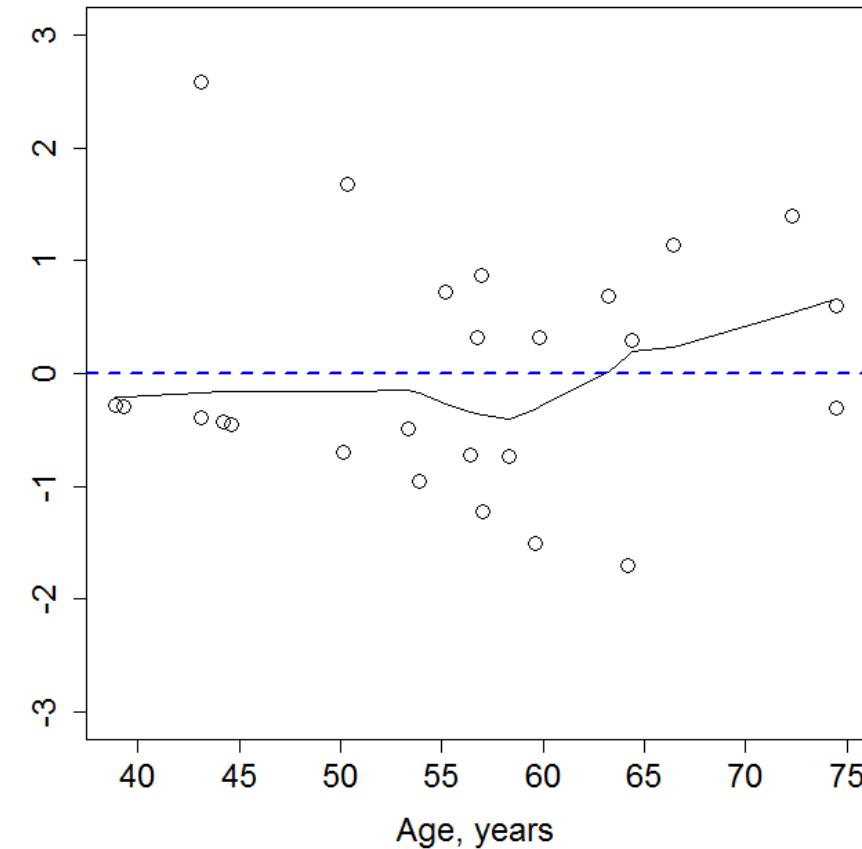
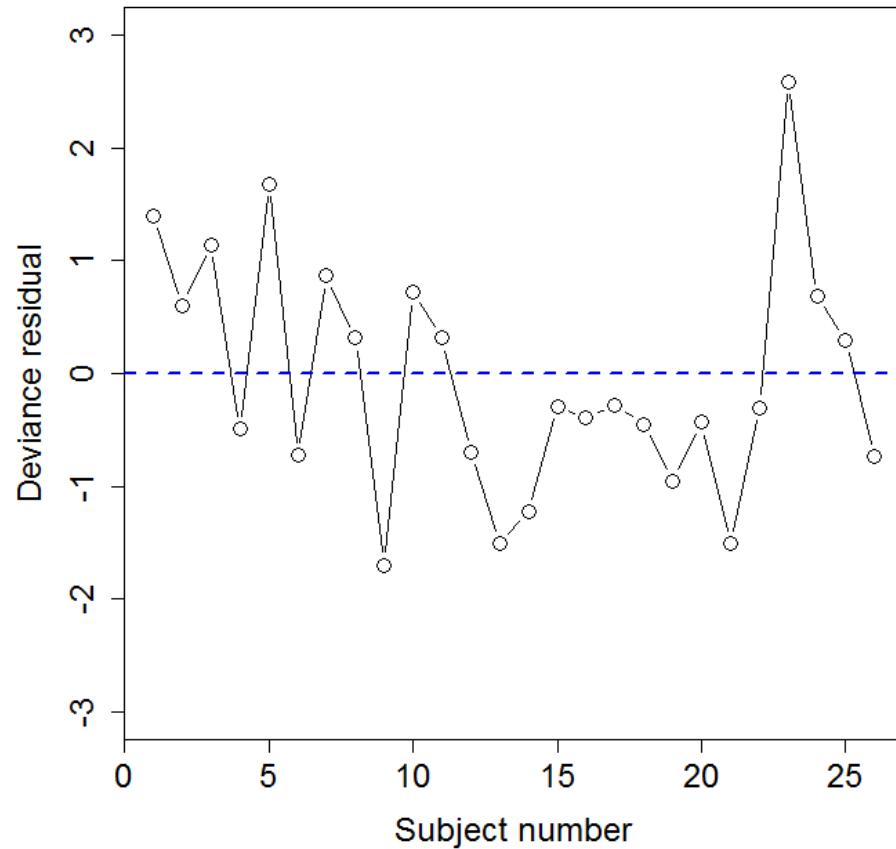
Martingale and Deviance Residuals



Martingale residual and functional form of covariate



Deviance residuals-Outlier detection



Summarize information from residuals

Type of residual	Feature	What kind o f plot?	What to look for?
Martingale (All data)	At each time point	Vs survival time	Skewed with mean zero, no inference
	At each time point	Vs predictor variable	Functional form of covariate
Deviance (All data)	At each time point	Vs Survival time	Random around zero, outlier detection
	At each time point	Vs predictor variable	Random around zero, Model adequacy
Schoenfeld (Only for event times)	For each predictor variable at their event times	Vs survival time	No trend upon smoothed fit, Proportional hazards assumption
Cox-snell	At each time point	Vs cumulative hazard of cox-snell residual	Random around the line of unity (45 degree line) : Overall fit of the model

What if proportional hazards assumption fails?

- Stratified Cox model can be used
 - Stratify the data based on a covariate
 - Different baseline hazards for each strata
 - Make inference within the strata

Time varying covariates

- Extended Cox Model

$$h(t; x) = h_o(t) \exp\left(\sum_{i=1}^p \beta_i x_i + \sum_{j=p+1}^k \beta_j x_j(t)\right)$$

- Hazard ratio is time varying and hence interpretation also becomes difficult

Exposure-Hazards relationship

Population pharmacokinetics and concentration–effect relationships of capecitabine metabolites in colorectal cancer patients

Ronald Gieschke, Hans-Ulrich Burger, Bruno Reigner,¹ Karen S. Blesch² & Jean-Louis Steimer

Biostatistics and ¹Clinical Pharmacology, Pharma Development, E Hoffmann-La Roche Ltd, Basel, Switzerland, and ²Biometrics, Pharma Development, EHoffmann-La Roche, Inc., Nutley, NJ, USA

2003 Blackwell Publishing Ltd *Br J Clin Pharmacol*, **55**, 252–263

sure to 5-FU was poorly predictive. The objective of the present analysis was to investigate further the relationships between systemic exposure (AUC and C_{max}) to the capecitabine metabolites 5'-DFUR, 5-FU, and FBAL and safety and efficacy outcomes in colorectal cancer patients. Values for AUC and C_{max} were derived from a population pharmacokinetic (PK) model constructed for this purpose.

Cox-proportional hazards model

The Cox proportional hazard model was used in the analysis of time-to-event data (i.e. time to disease progression and duration of survival). The underlying model for the hazard rate was

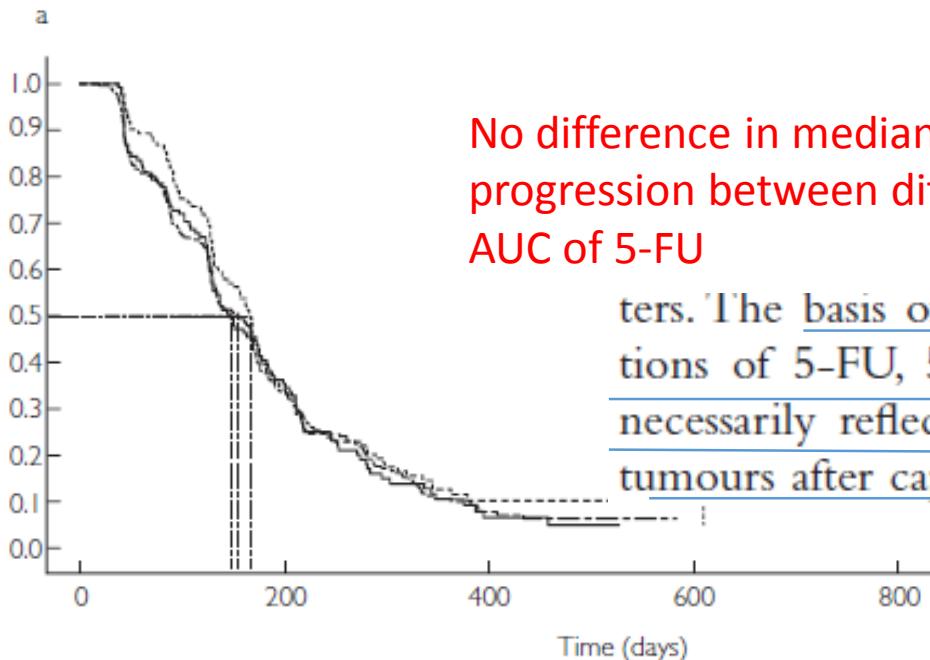
$$\lambda(t) = \lambda_0(t)\exp(-\beta^T Z)$$

Table 8 Cox regression analysis for systemic exposure and time to disease progression.

<i>Exposure measure</i>	<i>Hazard ratio</i>	<i>95% CI (Wald)</i>	<i>P-value (Wald)</i>
5'-DFUR AUC	0.996	0.97–1.02	0.75
5'-DFUR C_{max}	0.965	0.93–1.00	0.068
5-FU AUC	1.626	1.15–2.29	0.0056
5-FU C_{max}	1.097	0.72–1.67	0.66
FBAL AUC	0.989	0.97–1.01	0.16
FBAL C_{max}	0.918	0.84–1.01	0.063

n = 481, 426 events.

Clinical relevance of the finding



No difference in median time to disease progression between different quartiles of AUC of 5-FU

ters. The basis of these findings may be that concentrations of 5-FU, 5-DFUR and FBAL in plasma do not necessarily reflect concentrations in healthy tissues and tumours after capecitabine therapy. Our data suggest that

Figure 4 (a) Estimated probability for time to disease progression classified by the AUC of 5-FU (----, 1st; - - - -, 2nd and 3rd; —, 4th quartile). (b) Estimated probability for duration of

Simulating survival outcomes

- Parametric model to simulate survival times
- Accelerated failure time (AFT) model
- Natural logarithm of survival times is modeled as a function of covariates
 - AFT model postulates a direct relationship between failure time and covariates.

$$Y = \log(T) = \alpha + \beta' X + \sigma \epsilon$$

Residual error

Linear function
of predictors

Scale parameter

The diagram illustrates the components of the AFT model equation. The term $\beta' X$ is highlighted with a red box. Arrows point from the term $\sigma \epsilon$ to the label "Residual error" and from the term $\beta' X$ to the label "Linear function of predictors". A third arrow points from the term $\sigma \epsilon$ to the label "Scale parameter".

Distribution of survival times

Residual Error	Survival times
Normal distribution	Log-normal distribution
Extreme value distribution (2-parameter)	Weibull distribution
Extreme value distribution (1-parameter)	Exponential distribution
Logistic distribution	Log-logistic distribution

Scale parameter is sometimes omitted, fixed to one (exponential) or estimated (Weibull, log-normal)

Simulate survival times

- Simulate survival times for a drug known to be dependent on AUC of the drug and gender

Input model	
CL, AUC	CL ~ LN(7.5 L/h, 25%)
Covariate distribution model	
Gender	1:1 (males and females)
Trial Execution model	
Number of subjects	200
Censored	30%
Output model	
$T = \exp(0.02 + 0.05 \times AUC + 0.5 \times is.F + \epsilon)$	
$\epsilon \sim N(0, 0.05)$	

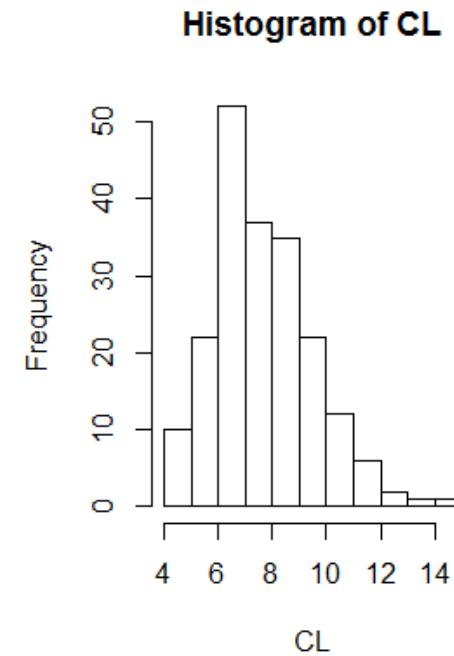
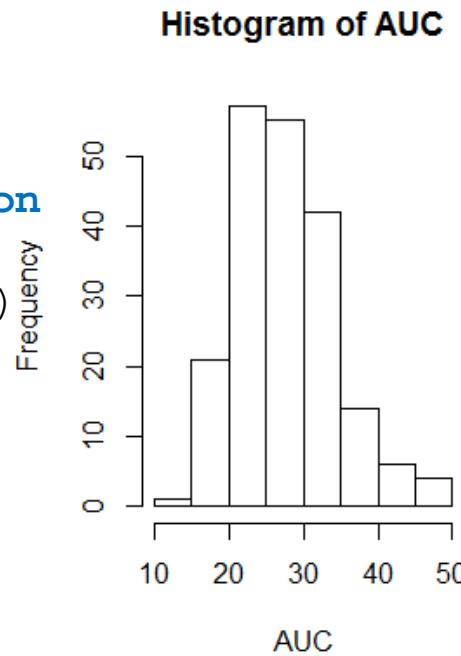
```

> ##### Using accelerated failure time models-----
> ##### Clearance of the drug and derive AUC-----
> n<-200
> Dose=200
> logCL<-rnorm(n,log(7.5),sd=0.25)
> CL<-exp(logCL)
> AUC<-Dose/CL
> par(mfrow=c(1,2))
> hist(AUC)
> hist(CL)

> ##### Covariate distribution
#####model-----
> is.female<-rbinom(n, 1, 0.5)
> table(is.female)
is.female
 0   1
104 96

> ##### Data frame to store
###survival times-----
>
> surv.data<-
data.frame(ID=1:n,AUC=AUC,
is.F=is.female)

```

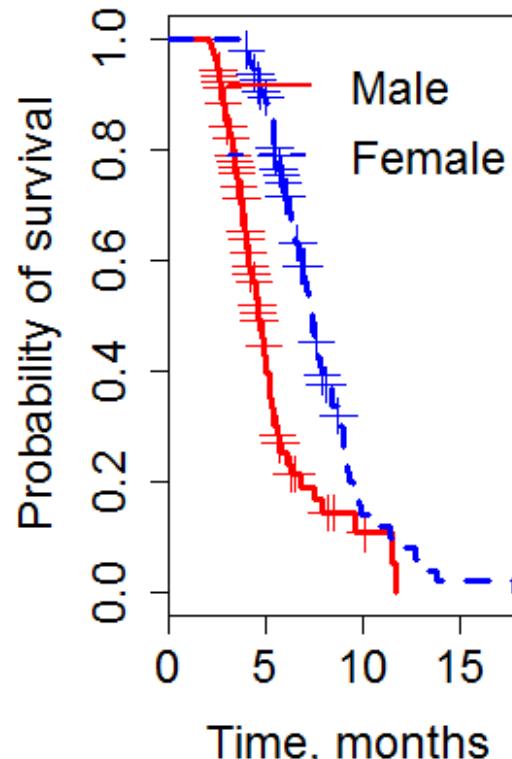
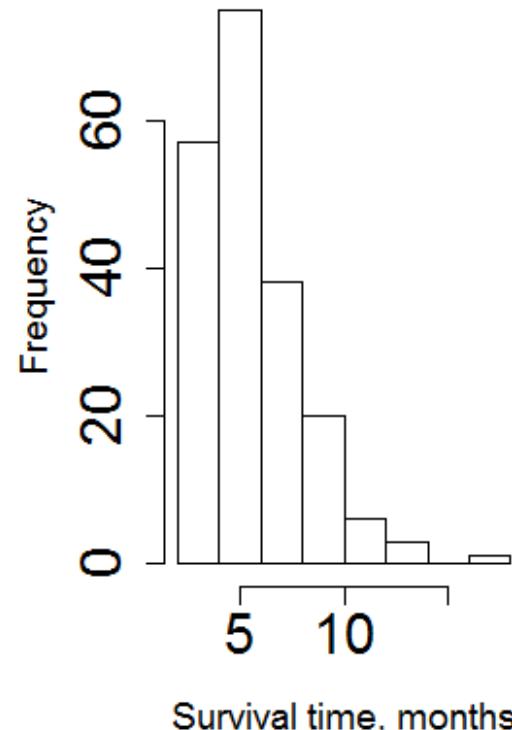


```

> #####SIMulate survival times using an AFT model-----
> eps<-rnorm(n,0,sqrt(0.0025))
> surv.data$TIME= exp(0.02+0.05*surv.data$AUC+0.5*surv.data$is.F+eps)
>
##### Censoring distribution: 30% censoring-----
> cens<-runif(n,0,1)
> surv.data$STATUS<-ifelse(cens<0.3,0,1)
> table(surv.data$STATUS)

```

0 1
65 135



Lunch

Exposure/Response Modeling – Dose Determination

- A pharmaceutical company is developing a new analgesic to be used for post-operative pain. A dose-ranging (phase I) study has been conducted in 160 patients, and you have been asked to provide input on the dose to be carried forward to phase II trials and beyond
- Drug tested at 3 dose levels (5 mg QD, 20 mg QD, and 80 mg QD)
- Placebo controlled trial
- Have drug conc vs time data (0-8hr post dose)
- Longitudinal binary response (pain relief: no=0; yes=1)
 - Response at each time point that PKs were taken

Primary Objective

- Select the optimal dose of drug for future clinical trials to give the drug the best chance to demonstrate significant efficacy with manageable off-target (side) effects

Final Dose Selection

- Apparent dose-proportional increases in drug exposure
- Apparent less-than dose-proportional increases in pain relief
- Marginalized models (using GEE) suggest “average” person would have lower odds of pain relief at 80mg vs 20mg
- GLMM suggests a subject at 80mg would have lower odds of pain relief (vs same subject on placebo) than a subject at 20mg (vs same subject on placebo)
- Final recommendation: 20 mg

Remedication

- Additional data available on our example case study from yesterday
- Patients that did NOT experience pain relief were offered drug (remedication)
- Time to remedication is a censorable statistic
- Can perform “survival” analyses, Cox model

Remedication

- Read in dataset

```
remed<-read.csv("TimetoRemed.csv", sep=",")  
head(remed)
```

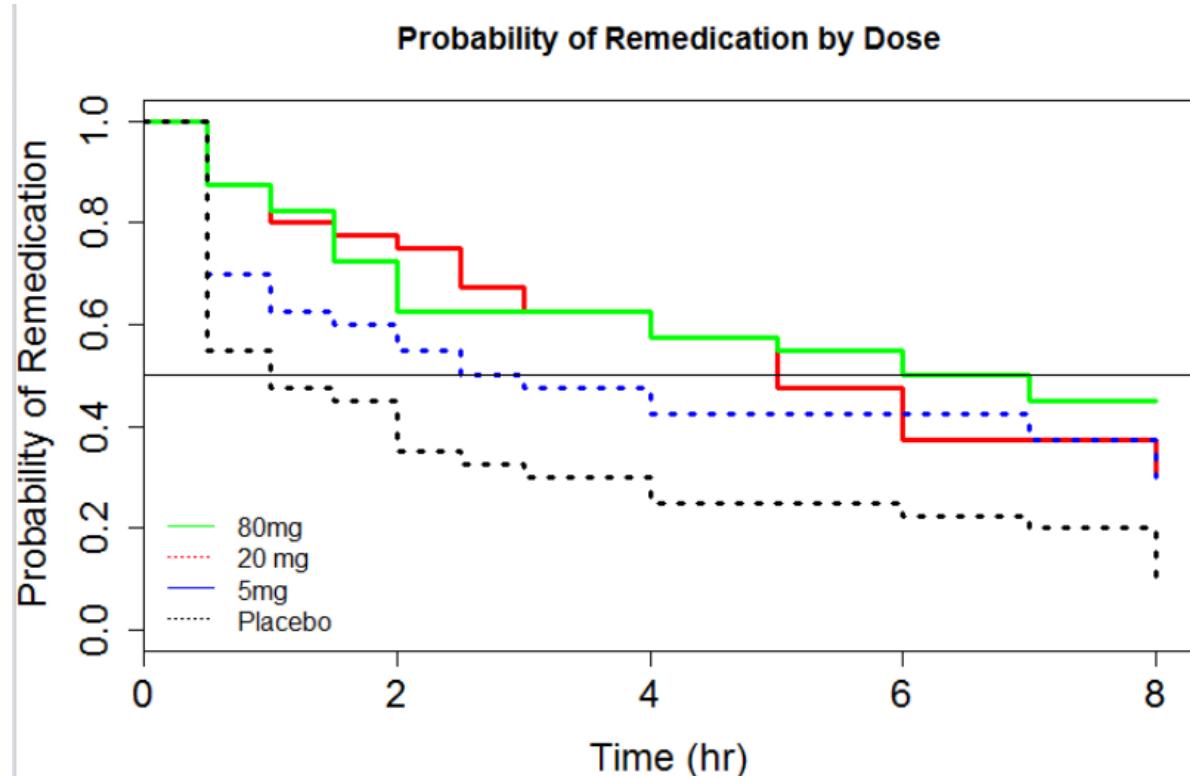
	ID	REMEDTime	REMEDStatus	ARM
1	1	8	0	A20_0_at2h
2	2	8	0	A80_0_at2h
3	3	8	0	Placebo
4	4	8	0	Placebo
5	5	8	0	A80_0_at2h
6	6	7	1	A5_0_at2h

- “REMEDTime” is the hour time point post initial dose where the patient took another dose of drug
- “REMEDStatus” is the censor status
 - 0 indicates remedication occurred at 8hr post first dose (no remedication)
 - 1 indicates remedication prior to 8 hr occurred
 - The hour post first dose where remedication occurred is listed in “REMEDTime”

Plot Remedication

- Performed survival fit
- 50% of patients given placebo asked for remedication within the first hour after taking placebo
- 50% of patients given 5 mg drug asked for remed just after 2hr post first dose
- 50% of patients given 20 mg drug asked for remed at 5 hr post first dose
- 50% of patients given 80 mg drug asked for remed at 6 hr post first dose
- First evidence of advantage to 80mg
 - Although not much separation between curves for 20mg and 80mg

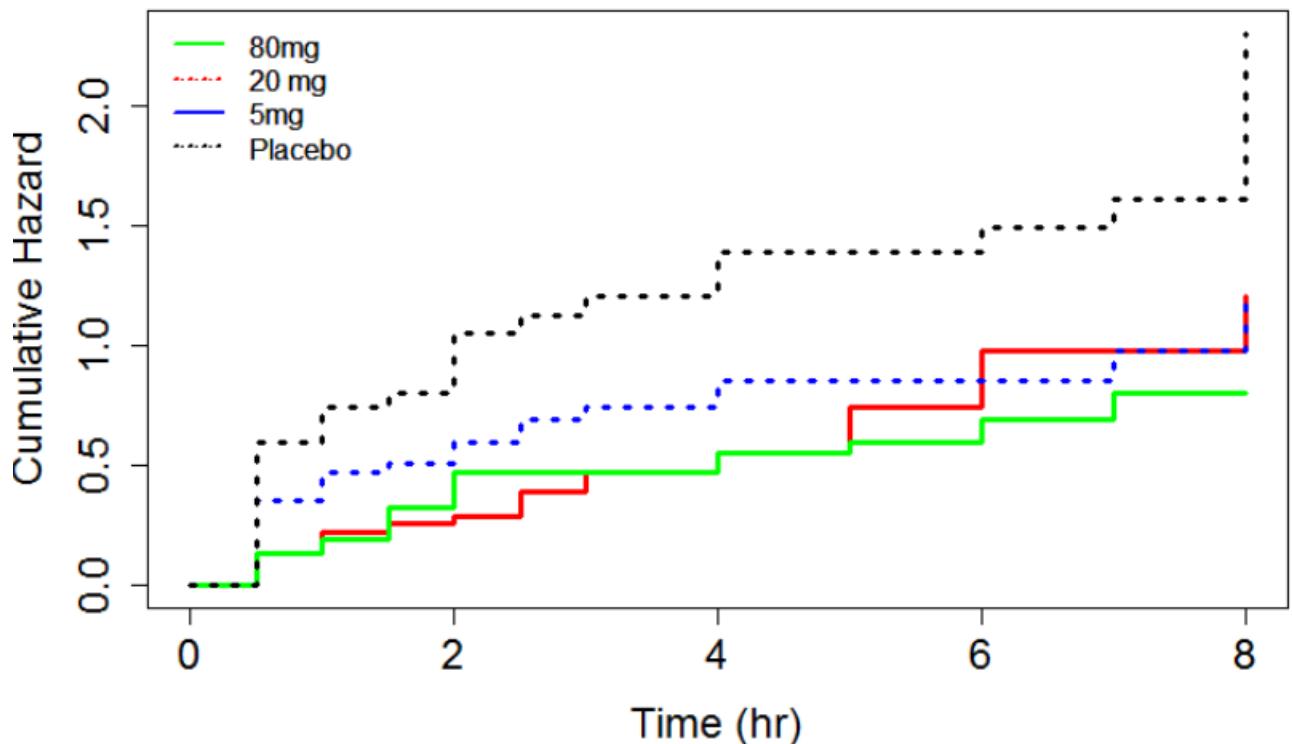
```
library(survival)
km.fit1<-survfit( Surv(REMEDTime, REMEDStatus)~ARM, data=remed)
summary(km.fit1)
plot(km.fit1,lty=c(1,3),cex=1.8,cex.lab=1.5,cex.axis=1.5,
     main="Probability of Remedication by Dose", xlab="Time (hr)",
     ylab="Probability of Remedication", lwd=3.0,
     col=c("red","blue", "green", "black"))
legend("bottomleft",c("80mg", "20 mg", "5mg", "Placebo"),lty=c(1,3),
       col=c("green","red", "blue", "black"),
       cex=1.0,lwd=1.0, bty="n")
abline(h=0.5)
```



Cumulative Hazards

- Assessing the cumulative hazard of each dose (0, 5, 20, 80mg) regarding the need to remediate
- Placebo group has highest cumulative hazard for event (remedication)
- 5mg and 20mg end up with similar hazards by 8hr
- 80mg has better separation from 20mg by 8hr, with observably lower cumulative hazard

```
#### Cumulative Hazard fnc; remed by dose---  
plot(km.fit1,fun="cumhaz",lty=c(1,3),cex=1.8,cex.lab=1.5,  
     cex.axis=1.5, xlab="Time (hr)",ylab="Cumulative Hazard",  
     lwd=3.0, col=c("red","blue", "green", "black"))  
legend("topleft",c("80mg","20 mg", "5mg", "Placebo"),lty=c(1,3),  
      col=c("green","red", "blue", "black"),  
      cex=1.0,lwd=2.0, bty="n")
```



Probability of Remedication

- To assess the likelihood/probability of a patient needing remedication, need to model the data
- Cox proportional hazards model
 - Calculates hazard ratio (HR): the probability of experiencing event within a specified time (given that they have not yet already experienced that event)
 - A subject that has already experienced event is considered “censored”
- Try full model first

Cox Model - FULL

- Bind AUC exposure data and Dose to the “remed” dataframe
- Make dose a factor

```
### bind AUC exposure data to remed data ##
nca<-read.csv("FinalPKNCA1.csv", header=TRUE)
remed2<-cbind(remed,nca$PPORRES_auclast)
remed3<-cbind(remed2, nca$DOSE)

remed4<-within(remed3,{DOSE<-factor(ARM,
  labels=c(0,5,20,80),levels=c("Placebo", "A5_0_at2h",
  "A20_0_at2h", "A80_0_at2h"))})
```

- Perform Cox model

Cox Model - FULL

- Bind AUC exposure data and Dose to the “remed” dataframe
- Make dose a factor

```
### bind AUC exposure data to remed data ##
nca<-read.csv("FinalPKNCA1.csv", header=TRUE)
remed2<-cbind(remed,nca$PPORRES_auclast)
remed3<-cbind(remed2, nca$DOSE)

remed4<-within(remed3,{DOSE<-factor(ARM,
  labels=c(0,5,20,80),levels=c("Placebo", "A5_0_at2h",
  "A20_0_at2h", "A80_0_at2h))})
```

- Perform Cox model

```
#### FULL Model ####
cph.full<- coxph(Surv(REMEDTime, REMEDStatus) ~ DOSE +
  nca$PPORRES_auclast, data = remed4, ties="breslow")
summary(cph.full)

Call:
coxph(formula = Surv(REMEDTime, REMEDStatus) ~ DOSE + nca$PPORRES_auclast,
       data = remed4, ties = "breslow")

n= 160, number of events= 114

            coef exp(coef) se(coef)      z Pr(>|z|)
DOSE5        -0.54112  0.58209  0.25909 -2.089  0.0367 *
DOSE20       -0.75471  0.47015  0.33817 -2.232  0.0256 *
DOSE80       -1.38390  0.25060  0.98663 -1.403  0.1607
nca$PPORRES_auclast  0.02217  1.02242  0.04341  0.511  0.6095
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

            exp(coef) exp(-coef) lower .95 upper .95
DOSE5          0.5821    1.7179   0.35031   0.9672
DOSE20         0.4701    2.1270   0.24232   0.9122
DOSE80         0.2506    3.9904   0.03624   1.7330
nca$PPORRES_auclast  1.0224    0.9781   0.93903   1.1132

Concordance= 0.614  (se = 0.034 )
Rsquare= 0.075  (max possible= 0.999 )
Likelihood ratio test= 12.45  on 4 df,  p=0.01433
Wald test           = 12.89  on 4 df,  p=0.01181
Score (logrank) test = 13.47  on 4 df,  p=0.009199
```

Cox Model - FULL

- 5mg and 20mg doses are significant predictors of remedication occurring
- 80mg is NOT
- AUC is NOT
 - Although technically, AUC and Dose are correlated, so should not include both in model together

```
#### FULL Model ####  
cph.full<- coxph(Surv(REMEDTime, REMEDStatus) ~ DOSE +  
  nca$PPORRES_auclast, data = remed4, ties="breslow")  
summary(cph.full)
```

```
Call:  
coxph(formula = Surv(REMEDTime, REMEDStatus) ~ DOSE + nca$PPORRES_auclast,  
      data = remed4, ties = "breslow")
```

n= 160, number of events= 114

	coef	exp(coef)	se(coef)	z	Pr(> z)
DOSE5	-0.54112	0.58209	0.25909	-2.089	0.0367 *
DOSE20	-0.75471	0.47015	0.33817	-2.232	0.0256 *
DOSE80	-1.38390	0.25060	0.98663	-1.403	0.1607
nca\$PPORRES_auclast	0.02217	1.02242	0.04341	0.511	0.6095

Signif. codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

	exp(coef)	exp(-coef)	lower .95	upper .95
DOSE5	0.5821	1.7179	0.35031	0.9672
DOSE20	0.4701	2.1270	0.24232	0.9122
DOSE80	0.2506	3.9904	0.03624	1.7330
nca\$PPORRES_auclast	1.0224	0.9781	0.93903	1.1132

Concordance= 0.614 (se = 0.034)
Rsquare= 0.075 (max possible= 0.999)
Likelihood ratio test= 12.45 on 4 df, p=0.01433
Wald test = 12.89 on 4 df, p=0.01181
Score (logrank) test = 13.47 on 4 df, p=0.009199

Cox Model - Reduced

- Remove insignificant covariates (AUC)
- Now, all dose levels are significant predictors
- Likely, the covariance of dose with AUC masked the predictive effects of 80mg dose
- Final model:

$$h(t) = h_0(t) \exp(\beta_1 \cdot DOSE)$$

$$h(t) = h_0(t) \exp(-0.513 \cdot 5mgDOSE)$$

$$h(t) = h_0(t) \exp(-0.641 \cdot 20mgDOSE)$$

$$h(t) = h_0(t) \exp(-0.905 \cdot 80mgDOSE)$$

```
#### reduced Model w/o AUC  
cph.red<- coxph(Surv(REMEDTime, REMEDStatus) ~ DOSE,  
                    data = remed4, ties="breslow")  
summary(cph.red)
```

Call:

```
coxph(formula = Surv(REMEDTime, REMEDStatus) ~ DOSE, data = remed4,  
      ties = "breslow")
```

n= 160, number of events= 114

	coef	exp(coef)	se(coef)	z	Pr(> z)
DOSE5	-0.5126	0.5989	0.2530	-2.026	0.042727 *
DOSE20	-0.6412	0.5267	0.2537	-2.527	0.011490 *
DOSE80	-0.9046	0.4047	0.2724	-3.321	0.000898 ***

Signif. codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

	exp(coef)	exp(-coef)	lower .95	upper .95
DOSE5	0.5989	1.670	0.3648	0.9833
DOSE20	0.5267	1.899	0.3203	0.8659
DOSE80	0.4047	2.471	0.2373	0.6903

Concordance= 0.616 (se = 0.033)

Rsquare= 0.073 (max possible= 0.999)

Likelihood ratio test= 12.19 on 3 df, p=0.006753

Wald test = 12.73 on 3 df, p=0.005255

Score (logrank) test = 13.28 on 3 df, p=0.004061

Cox Model - Reduced

$$h(t) = h_0(t) \exp(\beta_1 \cdot DOSE)$$

- **Exponentiated** coefficients are the hazard ratios (HR)
- Subjects on the 5mg dose have 0.589 (60%) of the hazard (40% less risk) than subjects on placebo
- Subjects on the 20mg dose have 0.527 (53%) of the hazard (47% less risk) than subjects on placebo

```
Call:  
coxph(formula = Surv(REMEDTime, REMEDStatus) ~ DOSE, data = remed4,  
      ties = "breslow")  
  
n= 160, number of events= 114  
  
          coef exp(coef) se(coef)     z Pr(>|z|)  
DOSE5   -0.5126    0.5989   0.2530 -2.026 0.042727 *  
DOSE20  -0.6412    0.5267   0.2537 -2.527 0.011490 *  
DOSE80  -0.9046    0.4047   0.2724 -3.321 0.000898 ***  
---  
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1  
  
          exp(coef) exp(-coef) lower .95 upper .95  
DOSE5    0.5989     1.670    0.3648    0.9833  
DOSE20   0.5267     1.899    0.3203    0.8659  
DOSE80   0.4047     2.471    0.2373    0.6903  
  
Concordance= 0.616  (se = 0.033 )  
Rsquare= 0.073  (max possible= 0.999 )  
Likelihood ratio test= 12.19  on 3 df,  p=0.006753  
Wald test           = 12.73  on 3 df,  p=0.005255  
Score (logrank) test = 13.28  on 3 df,  p=0.004061
```

- Subjects on the 80mg dose have 0.405 (41%) of the hazard (59% less risk) than subjects on placebo

Final Dose Selection

- Based on Cox model, 80mg shows a clear advantage over 20mg regarding time to remedication
- GEE and GLMM suggested 20mg as best dose choice
 - 80mg showed no additional pain relief benefit vs 20mg
 - Increasing exposure with 80mg may increase risk of side effects
- If 80mg better than 20mg for remedication, then make dosing schedule **20mg more frequently**
 - q4hr, q6hr, etc.
- OR, could recommend additional dose-finding studies to explore doses between 20 – 80 mg
- OR, could use 80mg *IF* the added exposure does NOT result in severe AEs

Conclusions

- Exposure/Response relationships can provide valuable insight into the status of your drug
 - Efficacy
 - Toxicity
- Can use to help make informed dosing decisions
- No matter what type of exposure or response data you have, the hope is that you can use the R codes presented here to apply to your datasets in the future

Conclusions

- Thank you for your attendance and attention
- Please fill out the accompanying survey to let us know how we can improve for next time