



Introduction to R Workshop

Amrom Obstfeld
August 16, 2020



Course Introduction



Goals and Objectives

- Advocate for the use of R as a means of improving reproducibility in clinical data analysis
- Demonstrate how R is used to perform analyses of laboratory operational data
- Establish a basis of understanding in the 'tidy' approach to data analysis within the framework of R



Who are we?



Stephan Kadauke

Assistant Professor of Clinical
Pathology and Laboratory Medicine

University of Pennsylvania Perelman
School of Medicine

Assistant Director of the Cell and
Gene Therapy Laboratory

Children's Hospital of Philadelphia



Daniel Herman

Assistant Professor of Pathology and
Laboratory Medicine

University of Pennsylvania Perelman
School of Medicine

Director, Endocrinology Laboratory

Hospital of the University of
Pennsylvania



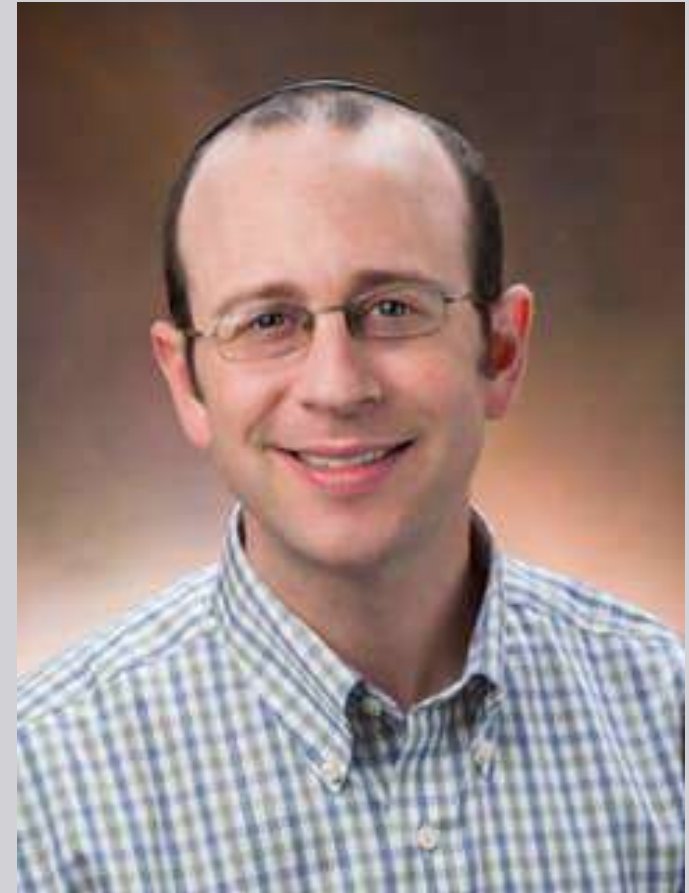
Amrom Obstfeld

Assistant Professor of Clinical
Pathology and Laboratory
Medicine

University of Pennsylvania
Perelman School of Medicine

Director of Pathology Informatics

Children's Hospital of
Philadelphia



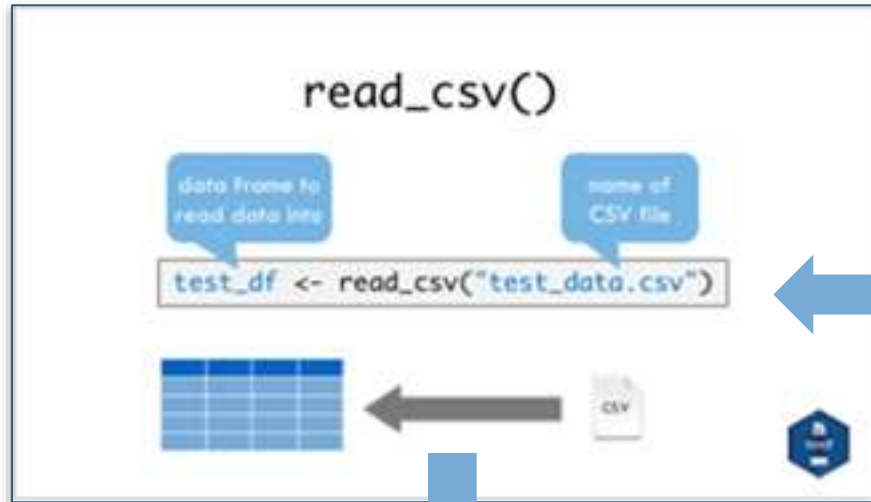


Workshop Workflow



August 16 2020	Session	Instructor
9:00 am - 9:30 am	Instructor Introductions, Introduction to technology	Amrom Obstfeld
9:30 am - 10:15 am	Introduction to R and RStudio	Amrom Obstfeld
10:30 pm - 11:15 am	Reproducible Reporting	Amrom Obstfeld
11:30 am - 1:00 am	Data Visualization	Stephan Kadauke
August 23 2020		
9:00 am - 10:30 pm	Data Transformation	Amrom Obstfeld
10:45 am - 12:15 pm	Statistical Analysis	Dan Herman
12:30 pm - 1:00 pm	Workshop Close out	Amrom Obstfeld

Sessions



Your Turn

Introduce yourself to your neighbors

- Who are you?
- Where are you from?
- What do you do with data?
- Have you ever used R?

3:00

```
1 ---
2 title: "R Notebook"
3 output: html_notebook
4 ---
5
6 This is an R Notebook. R Notebooks are written in R Markdown. An R
7 Notebook is like an electronic lab notebook, but for data analysis. You
8 can use R Notebooks to take notes, write code, and you can run that code
9 and see the results in the same document.
10
11 To take notes, simply edit the text in this document. For example, edit
12 the following line to replace XXX with your name:
13
14 My name is XXX, and I'm editing an R Notebook!
15
16 In an R Markdown document, code goes into "code chunks". Each code chunk
17 starts with three back-ticks (```) and the letter "r" in curly brackets.
18 It ends with a line that only has three backticks (```). The RStudio
19 editor makes the background color of code chunks gray. This way it's easy
20 to see where all the code chunks are. You can run the code in a code
21 chunk by clicking the green triangle in the upper right corner of the
22 code chunk. The results will appear beneath the chunk. Try it!
23
24 ```{r}
25 plot(cars)
26 ```
27
28 Good job!
29
30 You can open a new R Notebook by going to File > New File > R
31 Notebook.
```

Course Project

ESR Reference Range Study

Background and Objectives

One of our clinicians let us know that a lot of his patients, especially healthy elderly ones, receive ESR results flagged as abnormally high. He wonders whether our reference ranges are too stringent for this population.

The ESR is the measurement of the rate at which red blood cells (RBCs) settle in anticoagulated blood. RBCs settle faster when pro-inflammatory factors, such as ferritin, cause RBCs to stick together to form rouleaux. Pro-inflammatory factors tend to be concentrated in inflammatory states. Thus, the ESR is a nonspecific marker for inflammation.

It is known that a normal ESR value varies with age (higher in younger individuals) and gender (higher in females). However, there are multiple reference ranges in use by different clinical laboratories.

It is questionable whether inspection of a mixed population of ESR values (i.e., from healthy and sick individuals) allows classification of normal vs. abnormal values. This may be the case if there is clear separation of normal and abnormal values. In any case, it may be useful to visualize a sample of ESR values and highlight different options for cutoffs to better understand whether changing our reference range might be beneficial.

MGH

Age	Male	Female
all	<10 mm/h	<20 mm/h

ARUP Laboratories

Age	Male	Female
all	<10 mm/h	<20 mm/h

Quest Diagnostics

Age	Male	Female
<50	<10 mm/h	<20 mm/h
>50	<20 mm/h	<30 mm/h

LabCorp

Age	Male	Female
<50	<10 mm/h	<20 mm/h
>50	<20 mm/h	<30 mm/h

Bakerman's ABCs of Interpretive Laboratory Data

Age	Male	Female
Infants	<10 mm/h	<10 mm/h
Infants and children	0 - 10 mm/h	0 - 10 mm/h
<50	1 - 10 mm/h	1 - 10 mm/h
>50	Age 10 - 5	Age 10 - 5

Data Acquisition

Connect to the MGH Pathology Database (pathlab101.gemstone.org), and select the table: testresults, joined to be

Retrieve all ESR values from 2017, along with the following columns:

1. OutAge: Patient age at time of collection.
2. CollectDateTime: Date and time of collection.
3. PName: Patient name.
4. PNumber: Patient ID# (MGH).

5. PSex: Patient sex.
6. Result: ESR result value.
7. TestOrderName: Test order name.
8. UserGAFig: This column marks "high" and "low" results.

The ESR data was stored in the data frame: `esr`. Save the unsorted data frame on disk in case the database server goes down.

Downloaded from the database and clean up:

Data Exploration and Cleaning

This following is a column-by-column inspection and, if necessary, clean up of the data.

OutAge	CollectDateTime	PName	PNumber	PSex	Result
<col>	<col>	<col>	<col>	<col>	<col>
47	2017-01-01 10:00:00	JOHNSON, JAMES	1234567	F	5
55	2017-01-01 10:00:00	SMITH, JOHN	2345678	M	35
52	2017-01-01 10:00:00	WILSON, JANE	3456789	M	34
58	2017-01-01 10:00:00	DAVIS, JAMES	4567890	F	71
38	2017-01-01 10:00:00	WILSON, JANE	5678901	F	28
34	2017-01-01 10:00:00	JOHNSON, JAMES	6789012	F	5
44	2017-01-01 10:00:00	SMITH, JOHN	7890123	F	34
9	2017-01-01 10:00:00	DAVIS, JAMES	8901234	F	62
53	2017-01-01 10:00:00	WILSON, JANE	9012345	M	65
27	2017-01-01 10:00:00	SMITH, JOHN	0123456	M	70

1-10 of 48,573 rows (1-9 of 9 columns)

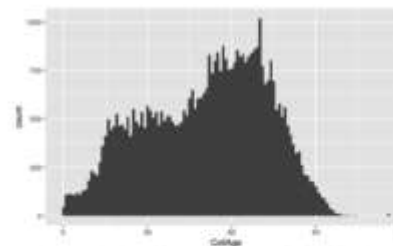
Missing values

Count the number of `NA`s (missing values) in each column of `esr`.

column	NA
<col>	<col>
OutAge	0
CollectDateTime	0
PName	0
PNumber	0
PSex	0
Result	0
TestOrderName	0
UserGAFig	0
Result	0

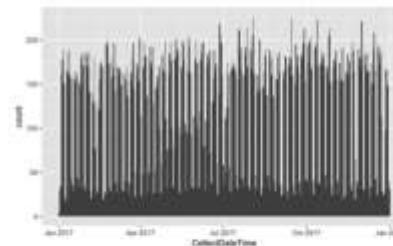
None of the columns of `esr` contain any missing values.

CollAge



`CollAge` - the age at collection, is an `integer` column, which is appropriate. The distribution of `CollAge` is bimodal, with peaks at around 30 and at around 70 years.

CollectDateTime



`CollectDateTime` is a `POSIXct` (timestamp) column, which is appropriate. Weekly cycling of ESR test volume is apparent. The weekly ESR test volume was approximately constant throughout 2017, without clear peaks or troughs.

PtName

```
[1] "JOHNSON, JAMES"
[2] "SMITH, JOHN"
[3] "WILSON, JANE"
[4] "DAVIS, JAMES"
[5] "WILSON, JANE"
[6] "JOHNSON, JAMES"
[7] "SMITH, JOHN"
[8] "DAVIS, JAMES"
[9] "WILSON, JANE"
[10] "SMITH, JOHN"
```

`PName` is a character column, which is appropriate. A random sample of 10 names is as expected.

PNumber

```
[1] "1234567"
[2] "2345678"
[3] "3456789"
[4] "4567890"
[5] "5678901"
[6] "6789012"
[7] "7890123"
[8] "8901234"
[9] "9012345"
[10] "0123456"
```

`PNumber` - the MGH IDs of the patients, is a character column, which is appropriate. Note: if `PNumber` were converted to integer, MGHs with leading zeros would be altered. A random sample of 10 MGHs shows that all are seven-digits long, as expected for MGH IDs.

PSex

`PSex` - sex character column, but since this is a categorical variable, we will convert it to a factor.

```
[1] F
[2] M
[3] F
[4] M
[5] F
[6] M
[7] F
[8] M
[9] F
[10] M
```

The majority of ESR tests were performed on female (77) patients. Unsurprisingly, in addition to the `PSex` column, we have a `Result` column.

OutAge	CollectDateTime	PName	PNumber	PSex	Result
<col>	<col>	<col>	<col>	<col>	<col>
55	2017-01-01 10:00:00	SMITH, JOHN	2345678	M	35
52	2017-01-01 10:00:00	WILSON, JANE	3456789	M	34
58	2017-01-01 10:00:00	DAVIS, JAMES	4567890	F	71
38	2017-01-01 10:00:00	WILSON, JANE	5678901	F	28
34	2017-01-01 10:00:00	JOHNSON, JAMES	6789012	F	5
44	2017-01-01 10:00:00	SMITH, JOHN	7890123	F	34
9	2017-01-01 10:00:00	DAVIS, JAMES	8901234	F	62
53	2017-01-01 10:00:00	WILSON, JANE	9012345	M	65
27	2017-01-01 10:00:00	SMITH, JOHN	0123456	M	70

1-10 of 17 rows

It appears that the rows with `Result` equals 0 belong to a small number of patients as well represent less than 0.1% of the whole data set, it will be able to remove them.

Result

`Result` is a character column but should be an integer, so we will convert it. Before doing so, we will

esr (converted by previous)

Result

```
<col>
```

```
12345
```

```
67890
```

```
12345
```

```
67890
```

```
12345
```

```
67890
```

```
12345
```

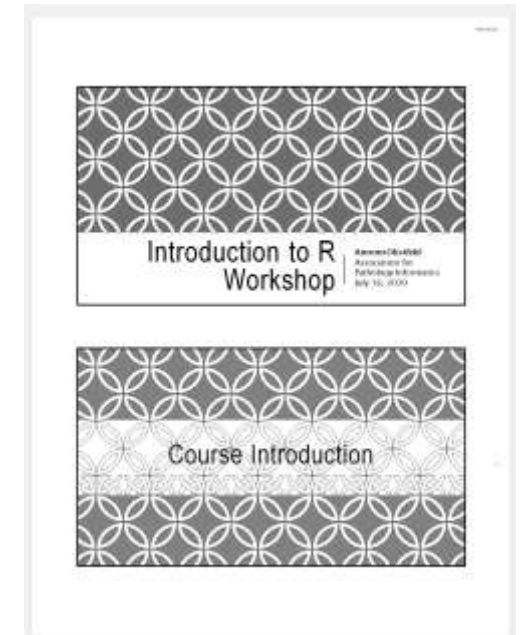
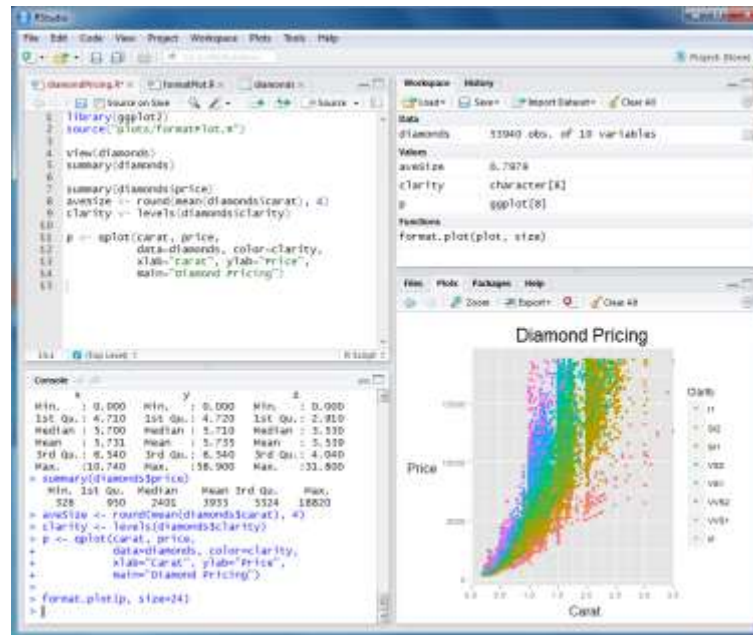
```
67890
```

Hackathons

- Participants work with experienced R users on the course project
- Helps troubleshoot problems
- Practice collaborating on analytics projects

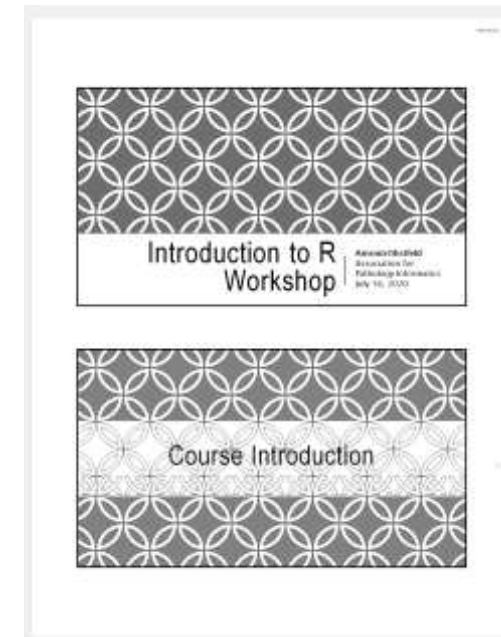


Your Setup



Workshop Coursebook

- Print out of all slides
- Appendix
 - Cheat sheets
 - Useful resources



Using Zoom



- Participants muted
- Chat window
- Non-verbal feedback
- Breakout sessions

Amrom

Using Zoom



- Participants muted
- Chat window
- Non-verbal feedback
- Breakout sessions



Unmute



Start Video



Participants



Chat



Share Screen



Record



Reactions

Leave

Amrom







Using Zoom









- Participants muted
- Chat window
- Non-verbal feedback
- Breakout sessions

Participants (2)

A Amrom (Me)  



NS Nova Smith (Host)  

raise Hand yes no go slower go faster more

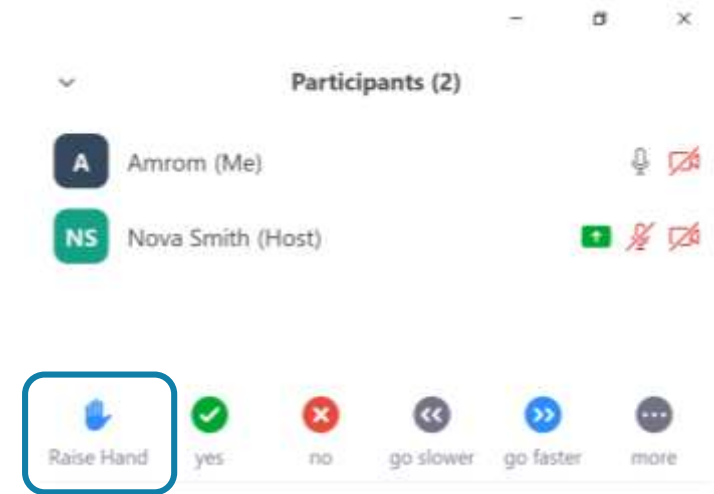
Invite Mute Me

Chat

To: Everyone ▾  File 

Type message here...

Getting Help



- During presentation – Raise hand, instructor will DM
- Break out sessions – Instructor available, unmuted



Who are you?

Your Turn

Introduce yourself!

Who are you?

Where are you from?

Why are you here?

Have you ever used R?

When you need more help

- The Internet (Stack Overflow:
<https://stackoverflow.com/>)
- Work Aids (RStudio Cheat Sheets:
<https://www.rstudio.com/resources/cheatsheets/>)
- A Good Book (R for Data Science:
<http://r4ds.had.co.nz/>)

Final Tips

- The best way to learn to code is by doing
- Practice is key!