Northwestern Medicine®

Enhancing Rigor and Transparency in Translational Research

Adopting Tools that Support Reproducible Research

Leah J. Welty, PhD
Director, Biostatistics Collaboration Center
Director, NUCATS Research Design Analysis Methods Program
Associate Professor
Department of Preventive Medicine, Division of Biostatistics
Department of Psychiatry and Behavioral Sciences

BCC: Biostatistics Collaboration Center

Who We Are

Leah J. Welty, PhD
Assoc. Professor
BCC Director

Joan S. Chmiel, PhD
Professor

Jody D. Ciolino, PhD
Asst. Professor

Kwang-Youn A. Kim, PhD
Asst. Professor

Masha Kocherginsky, PhD
Assoc. Professor

Mary J. Kwasny, ScD
Assoc. Professor

Julia Lee, PhD, MPH
Assoc. Professor

Alfred W. Rademaker, PhD
Professor

Hannah L. Palac, MS
Senior Stat. Analyst

Gerald W. Rouleau, MS
Stat. Analyst

Amy Yang, MS
Senior Stat. Analyst

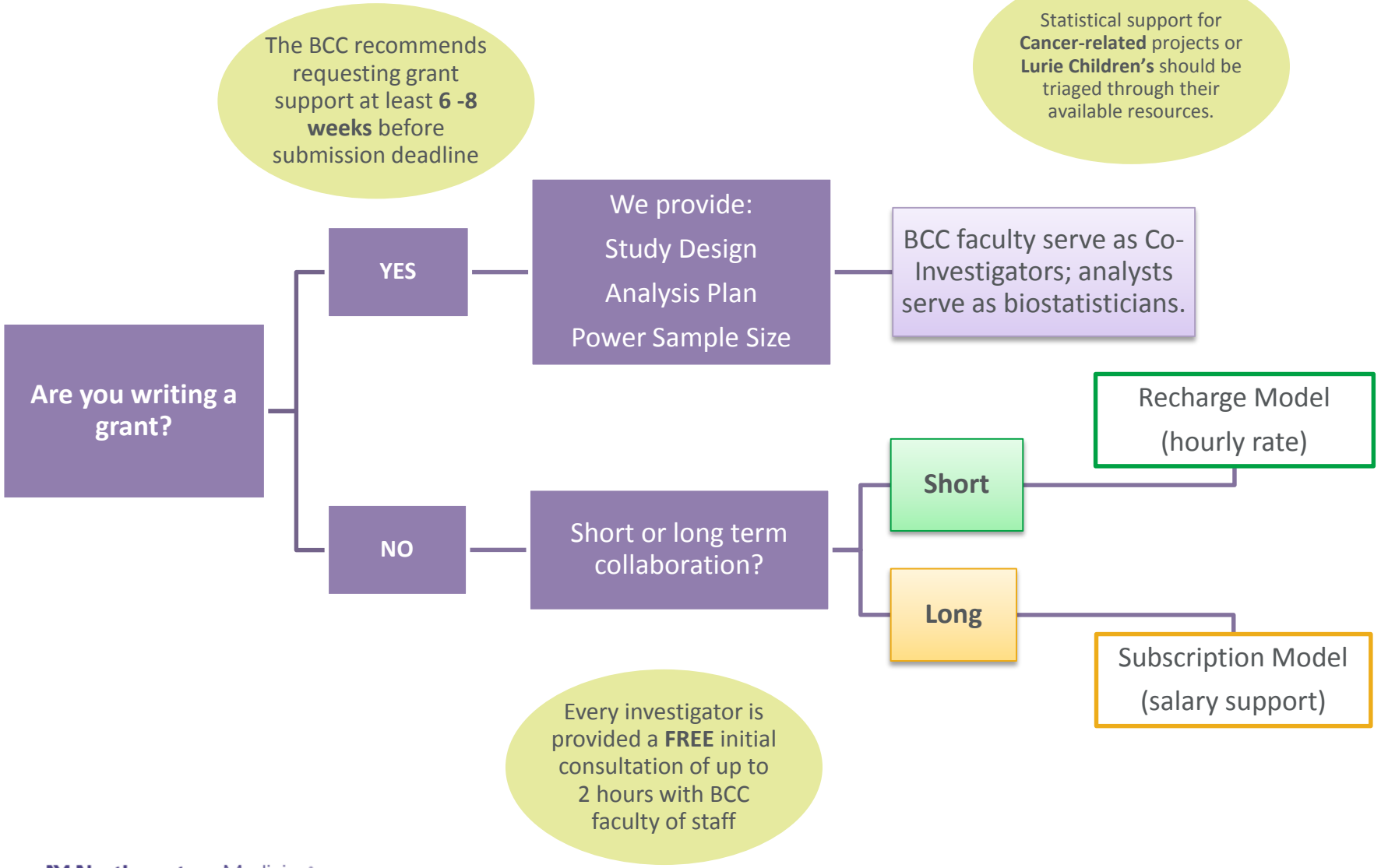
Not Pictured:
1. David A. Aaby, MS
Senior Stat. Analyst

Tameka L. Brannon
Financial | Research
Administrator

Northwestern Medicine®
Feinberg School of Medicine

BCC: Biostatistics Collaboration Center

How We Do It



BCC: Biostatistics Collaboration Center

Contact Us

- Request an Appointment
  - http://www.feinberg.northwestern.edu/sites/bcc/contact-us/request-form.html
- General Inquiries
  - bcc@northwestern.edu
  - 312.503.2288
- Visit Our Website
  - http://www.feinberg.northwestern.edu/sites/bcc/index.html

Biostatistics Collaboration Center |680 N. Lake Shore Drive, Suite 1400 |Chicago, IL 60611

Adopting Tools that Support Reproducible Research

Outline

- What is Reproducible Research?
- Why conduct Reproducible Research?
- Tools for Reproducible Research
  - Data Capture
  - Data Preparation and Analysis
  - Manuscript Preparation & StatTag
- Conclusions

Northwestern Medicine®

What is Reproducible Research?

What is Reproducible Research?

Reproducible vs Replicable

Reproducible vs Replicable

What is Reproducible Research?

Reproducible vs Replicable

Reproducible vs Replicable

Using the **same (raw) data** and information about the analysis methods and choices, we can recreate the results.

What is Reproducible Research?

Reproducible vs Replicable

Reproducible vs Replicable

Using the **same (raw) data** and information about the analysis methods and choices, we can recreate the results.

Scientific finding is verified or supported by **independent experiment or study**.

What is Reproducible Research?

Reproducible vs Replicable

Reproducible vs Replicable

Using the **same (raw) data** and information about the analysis methods and choices, we can recreate the results.

Scientific finding is verified or supported by **independent experiment or study**.

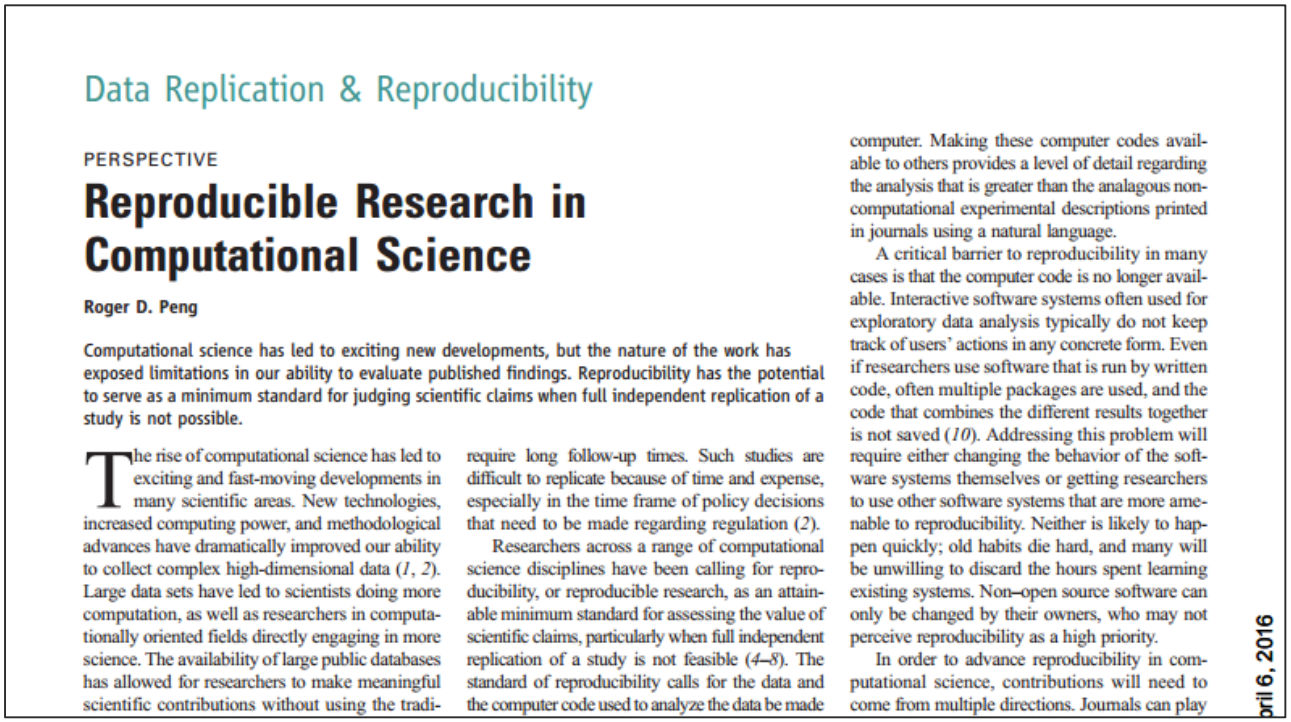
"We define reproducibility as the ability to re-compute data analytic results given an observed dataset and knowledge of the data analysis pipeline. The replicability of a study is the chance that an independent experiment targeting the same scientific question will produce a consistent result."

-- Leek and Peng

Leek and Peng "Opinion: Reproducible Research can still be wrong: Adopting a prevention approach" PNAS, February 10, 2015, vol. 112, no. 6, 1645–1646

What is Reproducible Research?

Strengths and Limitations



Data Replication & Reproducibility

PERSPECTIVE

Reproducible Research in Computational Science

Roger D. Peng

Computational science has led to exciting new developments, but the nature of the work has exposed limitations in our ability to evaluate published findings. Reproducibility has the potential to serve as a minimum standard for judging scientific claims when full independent replication of a study is not possible.

The rise of computational science has led to exciting and fast-moving developments in many scientific areas. New technologies, increased computing power, and methodological advances have dramatically improved our ability to collect complex high-dimensional data (*1*, *2*). Large data sets have led to scientists doing more computation, as well as researchers in computationally oriented fields directly engaging in more science. The availability of large public databases has allowed for researchers to make meaningful scientific contributions without using the traditional model of collecting one's own data, which can require long follow-up times. Such studies are difficult to replicate because of time and expense, especially in the time frame of policy decisions that need to be made regarding regulation (*2*).

Researchers across a range of computational science disciplines have been calling for reproducibility, or reproducible research, as an attainable minimum standard for assessing the value of scientific claims, particularly when full independent replication of a study is not feasible (*4*–*8*). The standard of reproducibility calls for the data and the computer code used to analyze the data be made available to others so that they may reproduce the findings in question. Making these computer codes available provides a level of detail regarding the analysis that is greater than the analagous non-computational experimental descriptions printed in journals using a natural language.

A critical barrier to reproducibility in many cases is that the computer code is no longer available. Interactive software systems often used for exploratory data analysis typically do not keep track of users' actions in any concrete form. Even if researchers use software that is run by written code, often multiple packages are used, and the code that combines the different results together is not saved (*10*). Addressing this problem will require either changing the behavior of the software systems themselves or getting researchers to use other software systems that are more amenable to reproducibility. Neither is likely to happen quickly; old habits die hard, and many will be unwilling to discard the hours spent learning existing systems. Non–open source software can only be changed by their owners, who may not perceive reproducibility as a high priority.

In order to advance reproducibility in computational science, contributions will need to come from multiple directions. Journals can play

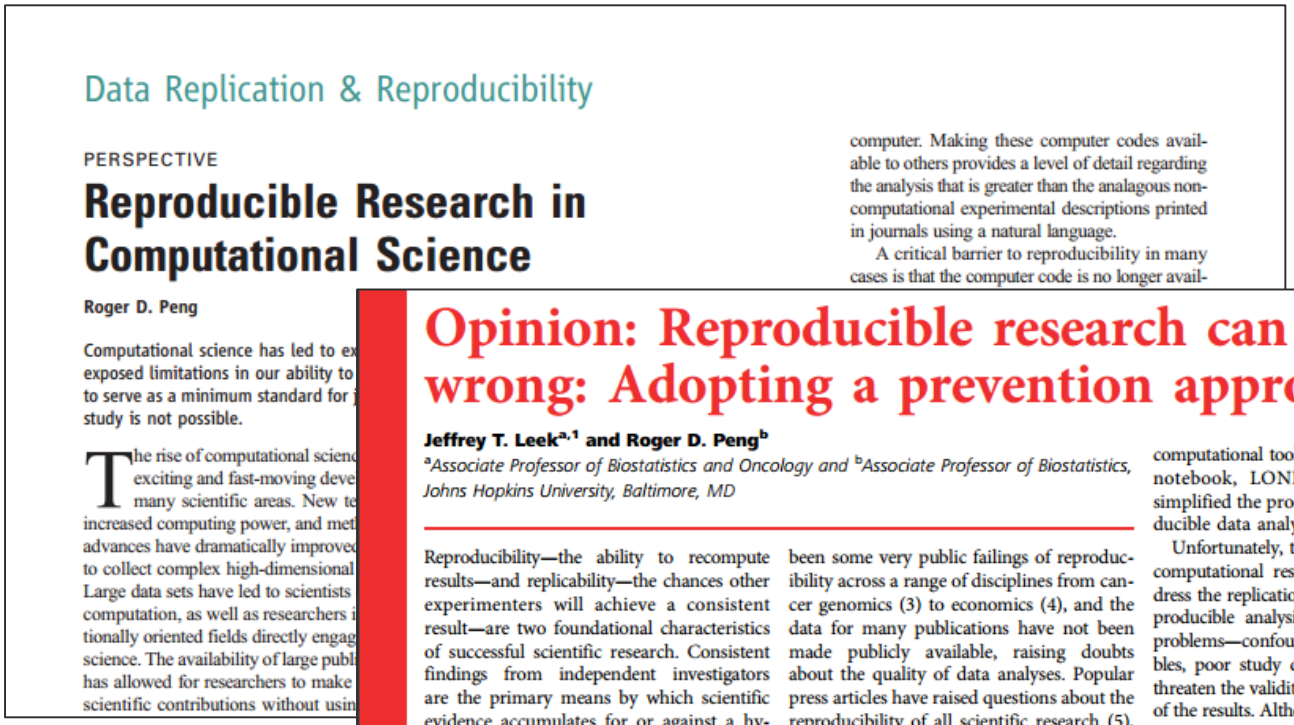
ril 6, 2016

Peng, R "Reproducible Research in Computational Science" (Science) 2 DECEMBER 2011 VOL 334

"Reproducibility has the potential to serve as a minimum standard for judging scientific claims when full independent replication of a study is not possible."

What is Reproducible Research?

Strengths and Limitations



Data Replication & Reproducibility

PERSPECTIVE

Reproducible Research in Computational Science

Roger D. Peng

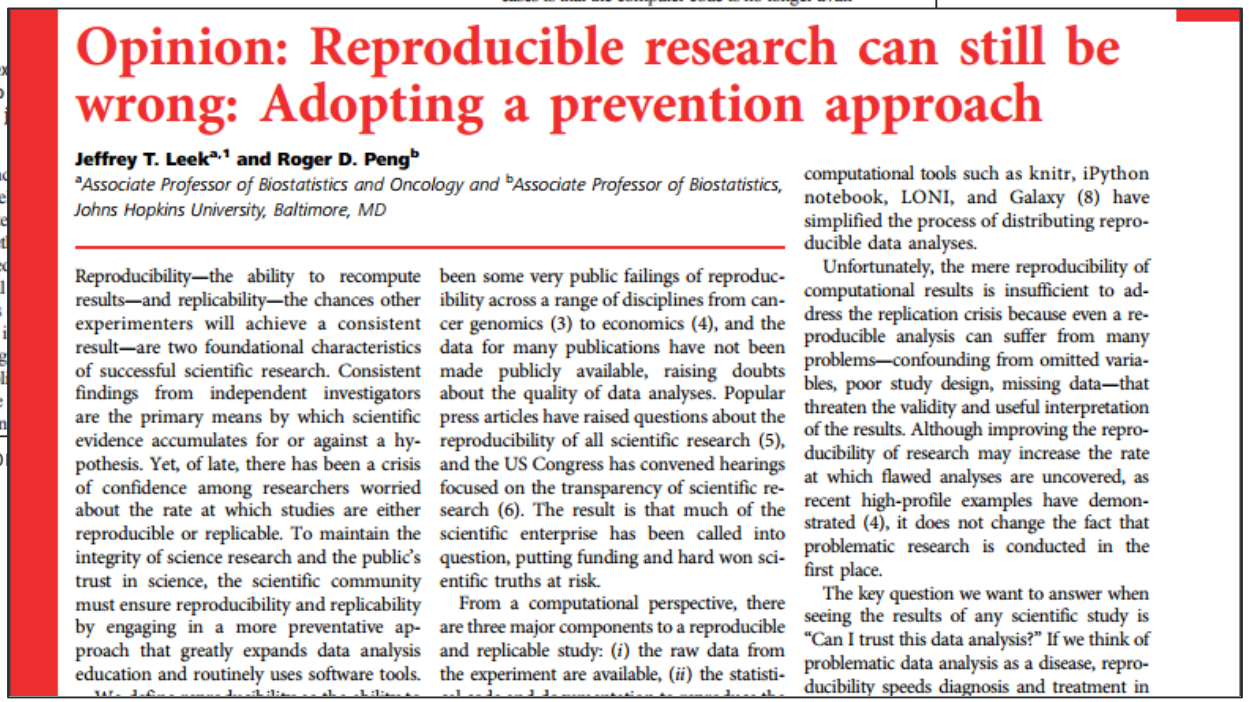
Computational science has led to ex
exposed limitations in our ability to
to serve as a minimum standard for
study is not possible.

The rise of computational scienc
exciting and fast-moving deve
many scientific areas. New te
increased computing power, and met
advances have dramatically improved
to collect complex high-dimensional
Large data sets have led to scientists
computation, as well as researchers i
tionally oriented fields directly engag
science. The availability of large publi
has allowed for researchers to make
scientific contributions without usin

computer. Making these computer codes available to others provides a level of detail regarding the analysis that is greater than the analagous non-computational experimental descriptions printed in journals using a natural language.

A critical barrier to reproducibility in many cases is that the computer code is no longer avail-

Peng, R "Rep



Opinion: Reproducible research can still be wrong: Adopting a prevention approach

Jeffrey T. Leek[a,1] and Roger D. Peng[b]

[a]Associate Professor of Biostatistics and Oncology and [b]Associate Professor of Biostatistics, Johns Hopkins University, Baltimore, MD

Reproducibility—the ability to recompute results—and replicability—the chances other experimenters will achieve a consistent result—are two foundational characteristics of successful scientific research. Consistent findings from independent investigators are the primary means by which scientific evidence accumulates for or against a hypothesis. Yet, of late, there has been a crisis of confidence among researchers worried about the rate at which studies are either reproducible or replicable. To maintain the integrity of science research and the public's trust in science, the scientific community must ensure reproducibility and replicability by engaging in a more preventative approach that greatly expands data analysis education and routinely uses software tools.

been some very public failings of reproducibility across a range of disciplines from cancer genomics (3) to economics (4), and the data for many publications have not been made publicly available, raising doubts about the quality of data analyses. Popular press articles have raised questions about the reproducibility of all scientific research (5), and the US Congress has convened hearings focused on the transparency of scientific research (6). The result is that much of the scientific enterprise has been called into question, putting funding and hard won scientific truths at risk.

From a computational perspective, there are three major components to a reproducible and replicable study: (*i*) the raw data from the experiment are available, (*ii*) the statisti-

computational tools such as knitr, iPython notebook, LONI, and Galaxy (8) have simplified the process of distributing reproducible data analyses.

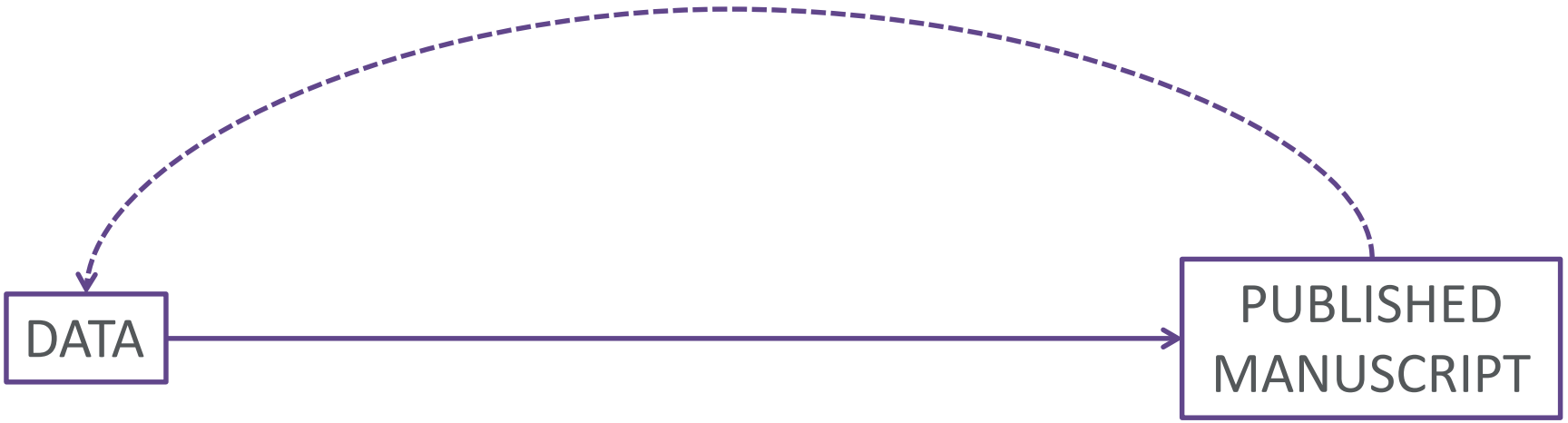
Unfortunately, the mere reproducibility of computational results is insufficient to address the replication crisis because even a reproducible analysis can suffer from many problems—confounding from omitted variables, poor study design, missing data—that threaten the validity and useful interpretation of the results. Although improving the reproducibility of research may increase the rate at which flawed analyses are uncovered, as recent high-profile examples have demonstrated (4), it does not change the fact that problematic research is conducted in the first place.

The key question we want to answer when seeing the results of any scientific study is "Can I trust this data analysis?" If we think of problematic data analysis as a disease, reproducibility speeds diagnosis and treatment in

Leek and Peng "Opinion: Reproducible Research can still be wrong: Adopting a prevention approach" PNAS, February 10, 2015, vol. 112, no. 6, 1645–1646

What is Reproducible Research?

What is Reproducible Research?



Northwestern Medicine®

Why conduct Reproducible Research?

Why conduct Reproducible Research?

Origins lie in the inconvenience of irreproducible research

"In our laboratory [Stanford Exploration Project in the 1980s] we noticed that after a few months or years, researchers were usually unable to reproduce their work without considerable agony."

-- Jon Claerbout

Why conduct Reproducible Research?

My own story

Data point error - in post-sertraline sham for pt 14

[redacted]

Sent: Sun 2/1/2015 4:57 PM

To: Leah J Welty; Amy Yang

Message | ALLsubjVALVAR5.xlsx (22 KB)

Leah,

I'm horrified to find a transcription error for one of the % total variance values, and it affects a single data point.
It affects some analyses, but no
So, that's lucky.

I was preparing data for copyin
I reviewed screenshots of the F
So far, I have found one error.

For patient 14 the post-sertralin
The % total variance was origina

The correct % total variance for
The corrected z*ptv value for p

I have attached a corrected exc

My apologies!

Leah,

I'm horrified to find a transcription error for one of the values, it affects a single data point … This is in spite of multiple layers of redundant double-checking when I first compiled the data. It's amazing to me that it got through….The corrected value for patient 14 for the sham test is 3.15 (instead of 0.47). …I have attached the corrected excel file.

My apologies!

Why conduct Reproducible Research?

My own story

Data point error - in post-sertraline sham for pt 14

[redacted]

Sent: Sun 2/1/2015 4:57 PM

To: Leah J Welty; Amy Yang

Message | ALLsubjVALVAR5.xlsx (22 KB)

Leah,

I'm horrified to find a transcription error for one of the % total variance values, and it affects a single data point.
It affects some analyses, but no
So, that's lucky.

I was preparing data for copying
I reviewed screenshots of the F
So far, I have found one error.

For patient 14 the post-sertralin
The % total variance was origina

The correct % total variance for
The corrected z*ptv value for p

I have attached a corrected exc

My apologies!

Leah,

I'm horrified to find a transcription error for one of the values, it affects a single data point … This is in spite of multiple layers of redundant double-checking when I first compiled the data. It's amazing to me that it got through….The corrected value for patient 14 for the sham test is 3.15 (instead of 0.47). …I have attached the corrected excel file.

My apologies!

Why conduct Reproducible Research?

My own story

Data point error - in post-sertraline sham for pt 14

[redacted]

Sent: Sun 2/1/2015 4:57 PM

To: Leah J Welty; Amy Yang

Message | ALLsubjVALVAR5.xlsx (22 KB)

Leah,

I'm horrified to find a transcription error for one of the % total variance values, and it affects a single data point.
It affects some analyses, but no[t] [redacted]
So, that's lucky.

I was preparing data for copying
I reviewed screenshots of the F
So far, I have found one error.

For patient 14 the post-sertralin
The % total variance was origina

The correct % total variance for
The corrected z*ptv value for p

I have attached a corrected exc

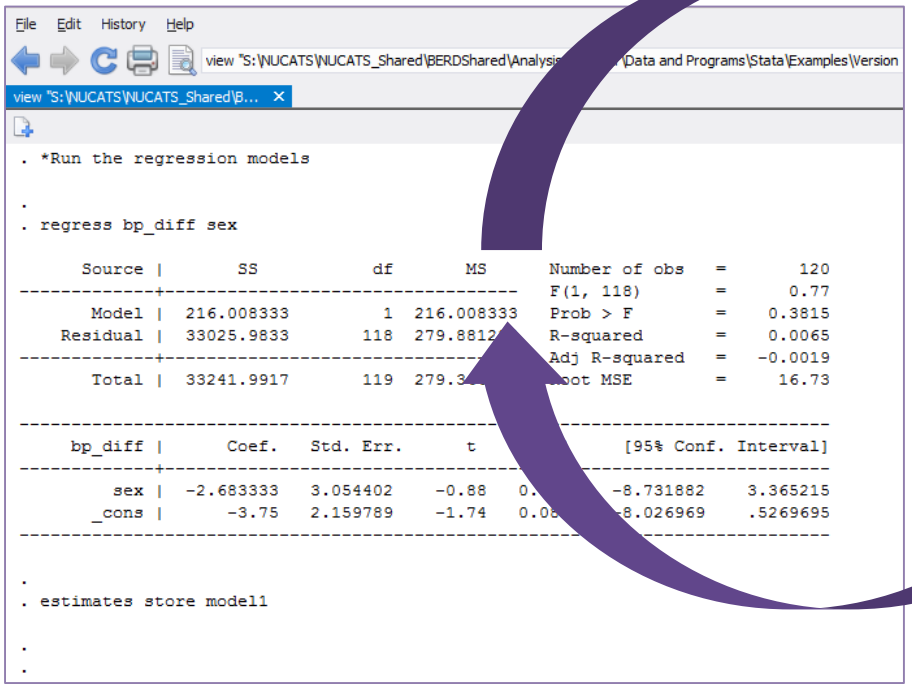
My apologies!

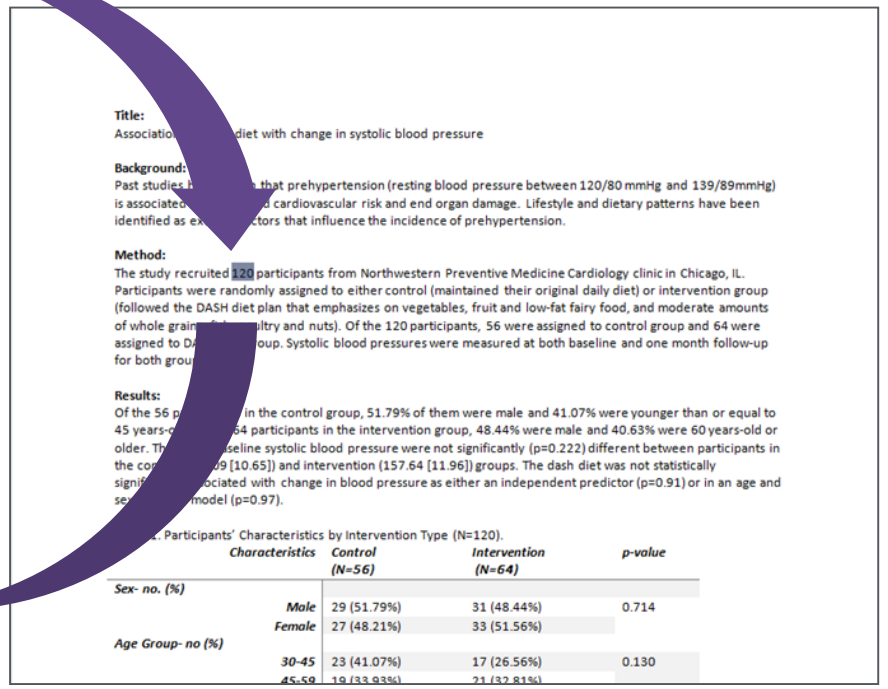
Leah,

I'm horrified to find a transcription error for one of the values, it affects a single data point … This is in spite of multiple layers of redundant double-checking when I first compiled the data. It's amazing to me that it got through….The corrected value for patient 14 for the sham test is 3.15 (instead of 0.47). …I have attached the corrected excel file.

My apologies!

Problem: Reproducible Research




Title:
Associatio diet with change in systolic blood pressure

Background:
Past studies h that prehypertension (resting blood pressure between 120/80 mmHg and 139/89mmHg) is associated d cardiovascular risk and end organ damage. Lifestyle and dietary patterns have been identified as ex ctors that influence the incidence of prehypertension.

Method:
The study recruited 120 participants from Northwestern Preventive Medicine Cardiology clinic in Chicago, IL. Participants were randomly assigned to either control (maintained their original daily diet) or intervention group (followed the DASH diet plan that emphasizes on vegetables, fruit and low-fat fairy food, and moderate amounts of whole grain ultry and nuts). Of the 120 participants, 56 were assigned to control group and 64 were assigned to DA oup. Systolic blood pressures were measured at both baseline and one month follow-up for both grou

Results:
Of the 56 p in the control group, 51.79% of them were male and 41.07% were younger than or equal to 45 years-o 64 participants in the intervention group, 48.44% were male and 40.63% were 60 years-old or older. Th seline systolic blood pressure were not significantly (p=0.222) different between participants in the co 09 [10.65]) and intervention (157.64 [11.96]) groups. The dash diet was not statistically signif ociated with change in blood pressure as either an independent predictor (p=0.91) or in an age and se model (p=0.97).

Participants' Characteristics by Intervention Type (N=120).

Characteristics	Control (N=56)	Intervention (N=64)	p-value
Sex- no. (%)			
Male	29 (51.79%)	31 (48.44%)	0.714
Female	27 (48.21%)	33 (51.56%)	
Age Group- no (%)			
30-45	23 (41.07%)	17 (26.56%)	0.130
45-59	19 (33.93%)	21 (32.81%)	

Have you ever had to:

- Copy results from statistical output to MS Word?
- Re-copy results when data or analyses change?
- Wondered some time after publication how you obtained an estimate?

Northwestern Medicine®

Why conduct Reproducible Research?

Avoid the copy paste nightmare

lifetime_linreg - Notepad

File Edit Format View Help

```
      name:  <unnamed>
       log:  P:\Products\Papers\Kid's Papers\Kid's Dev SUD\Analysis\log\lifetim
  log type:  text
 opened on:  24 Apr 2013, 13:09:07


. *sedative use

. qui wt_corr o_seddsm_lf
0.22%
.
. svy, sub(if male==1 & raceself !=4): logistic o_seddsm_lf black hisp
(running logistic on estimation sample)

Survey: Logistic regression

Number of strata   =         9                  Number of obs      =      1165
Number of PSUs     =      1165                  Population size    = 1684.3324
                                                 Subpop. no. of obs =      1165
                                                 Subpop. size       = 1684.3324
                                                 Design df          =      1156


------------------------------------------------
             |                 Linearized
 o_seddsm_lf | Odds Ratio   Std. Err.
-------------+----------------------------------
       black |   .0030322   .0031084    -5.
        hisp |   .1025896   .0535013    -4.
       _cons |   .1309232   .0292753    -9.
------------------------------------------------
Note: 4 strata omitted because they contain

.
. svy, sub(if male==1 & raceself !=4): logi
(running logistic on estimation sample)
```

All new results:

Ctrl-C Ctrl-V

How many times?

Disorder During the 12 Years After Detention, Cook County (Chicago), 1995-2011[a]

	Cocaine Use Disorder		Hallucinogen/PCP Use Disorder		Opiate Use Disorder		Amphetamine Use Disorder[b]		Sedative Use Disorder[b]	
Variable	AOR[h]	95% CI	AOR[h]	95% CI	AOR[h]	95% CI	AOR[h]	95% CI	AOR[h]	95% CI
Main effects										
Sex										
F vs M	1.41	(0.98, 2.04)	1.24	(0.78, 1.97)	**2.43**	**(1.41, 4.17)**	**3.29**	**(1.40, 7.73)**	**2.68**	**(1.11, 6.47**
Race/ethnicity										
W vs AA	**32.11**	**(13.80, 74.72)**	**18.8**	**(9.0, 39.3)**	**55.94**	**(12.88, 242.9)**				
W vs H	**1.52**	**(1.04, 2.21)**	**2.47**	**(1.59, 3.84)**	**7.49**	**(3.97, 14.13)**	**20.97**	**(6.69, 65.77)**	**11.88**	**(4.64, 30.4**
H vs AA	**21.18**	**(8.96, 50.08)**	**7.62**	**(3.54, 16.4)**	**7.47**	**(1.55, 36.12)**				
Time[d]	1.05	(1.00, 1.10)	**0.85**	**(0.78, 0.92)**	**1.12**	**(1.01, 1.24)**	**0.84**	**(0.74, 0.95)**	0.90	(0.79, 1.03

Abbreviations: AOR = Adjusted Odds Ratio; CI = Confidence Interval; AA = African American; H = Hispanic; W = non-Hispanic white; M = Male; F

[a] Odds ratios and their associated 95% confidence intervals were estimated via generalized estimating equations (GEEs), with linear and qua since baseline. GEE models were weighted to account for sampling design, and were adjusted for age at baseline (centered at 16 years of a (processed in juvenile or adult court). Because incarceration may restrict access to substances, all models also include covariates for time in

Northwestern Medicine®

What is Reproducible Research?

Tools and Mindset

Reproducible Research exists along a spectrum.

It is about taking a considered approach to your data collection and analysis pipeline.

There are tools that can help. They are fast evolving. Adoption of some or all is an improvement.

What is Reproducible Research?

Tools and Mindset

Reproducible Research exists along a spectrum.

It is about taking a considered approach to your data collection and analysis pipeline.

There are tools that can help. They are fast evolving. Adoption of some or all is an improvement.

However,

The most important tool is the mindset, when starting, that the end product will be reproducible.

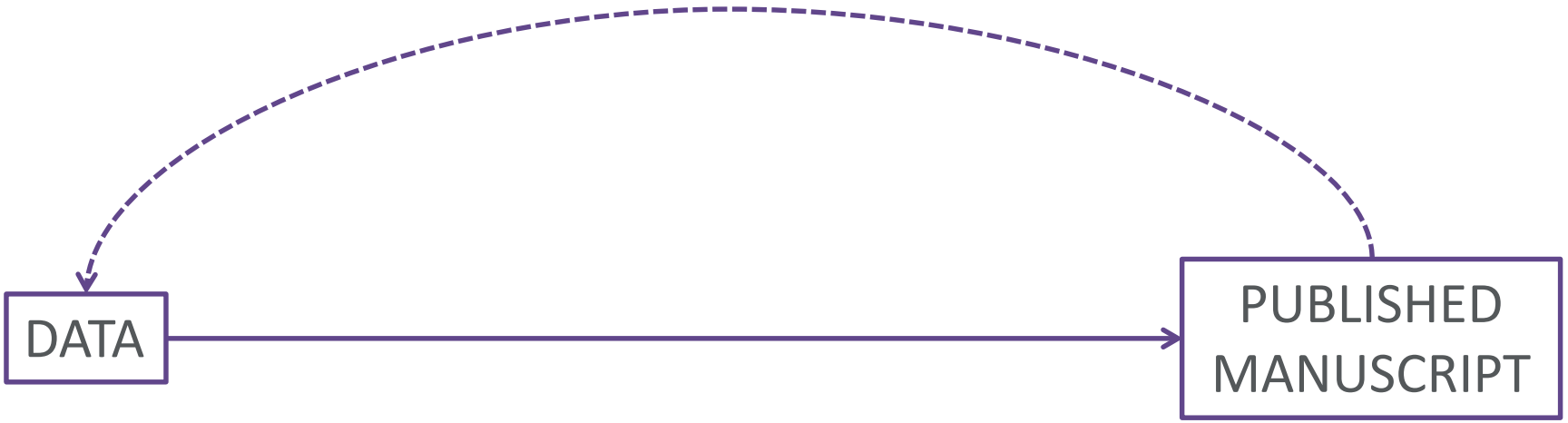
-- Keith Baggerly

Northwestern Medicine®

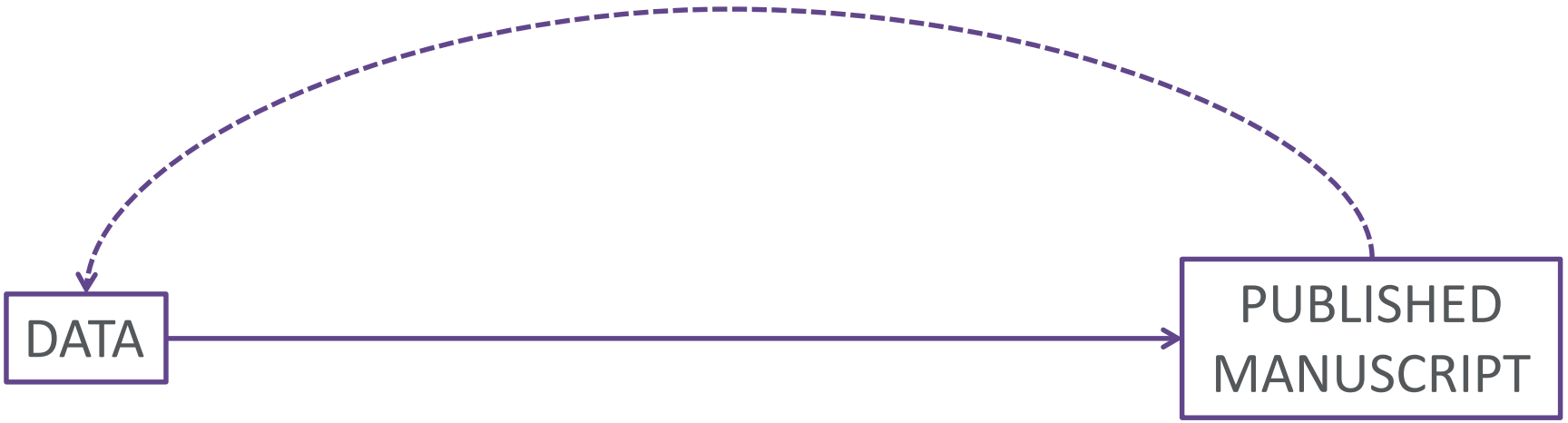
Tools for Conducting Reproducible Research

Data Capture
Data Preparation and Analysis
Manuscript Preparation

What is Reproducible Research?



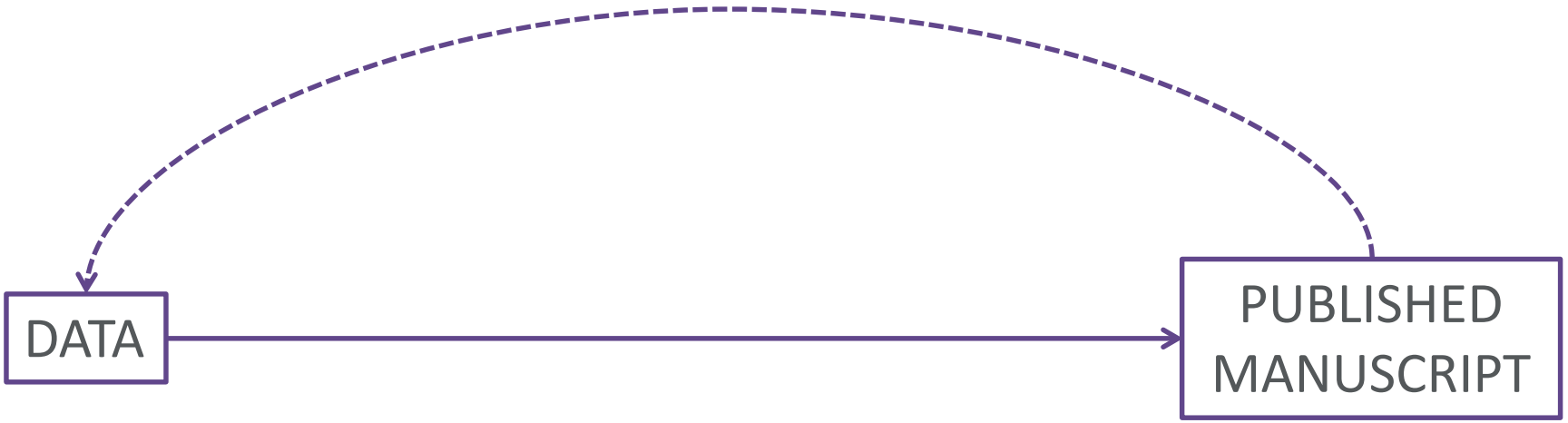
What is Reproducible Research?



Data capture --> Data preparation --> Data Analysis --> Interpretation

What is Reproducible Research?

Tools



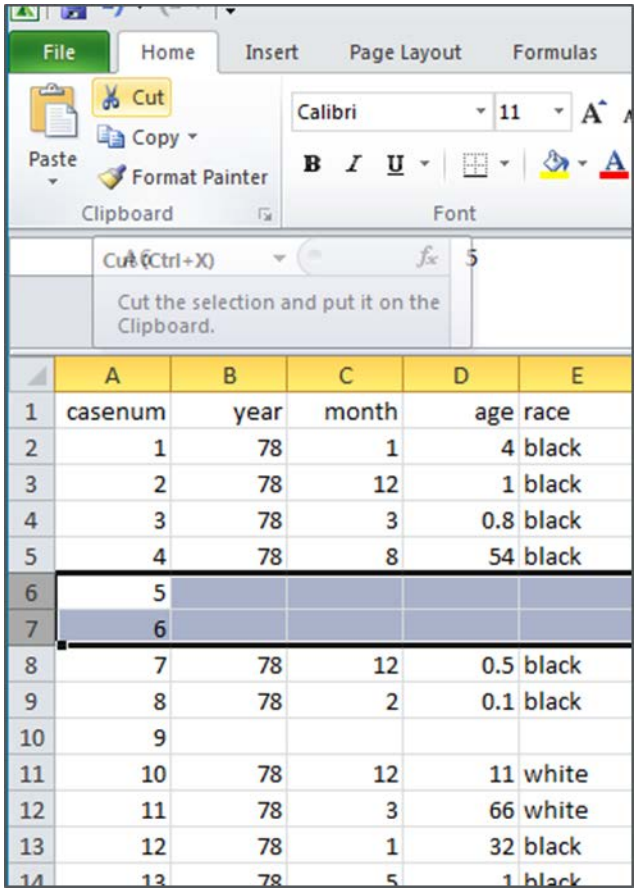
Data capture --> Data preparation --> Data Analysis --> Interpretation

What are tools that make each of these steps more reproducible?

Tools for Reproducible Research: Data Capture

What about Excel?

- Scenario
  - Going through EMR, interviewing patients, or taking measurements and entering data in to an Excel spreadsheet
- No explicit version control, trace-back, record or date stamp.
  - I know of people who lost entire studies this way.
- One-off/misalignment errors.
  - Wide spreadsheets
- Standardization
  - E.g. Black vs black
  - Inconsistent missing codes
  - Rounding values

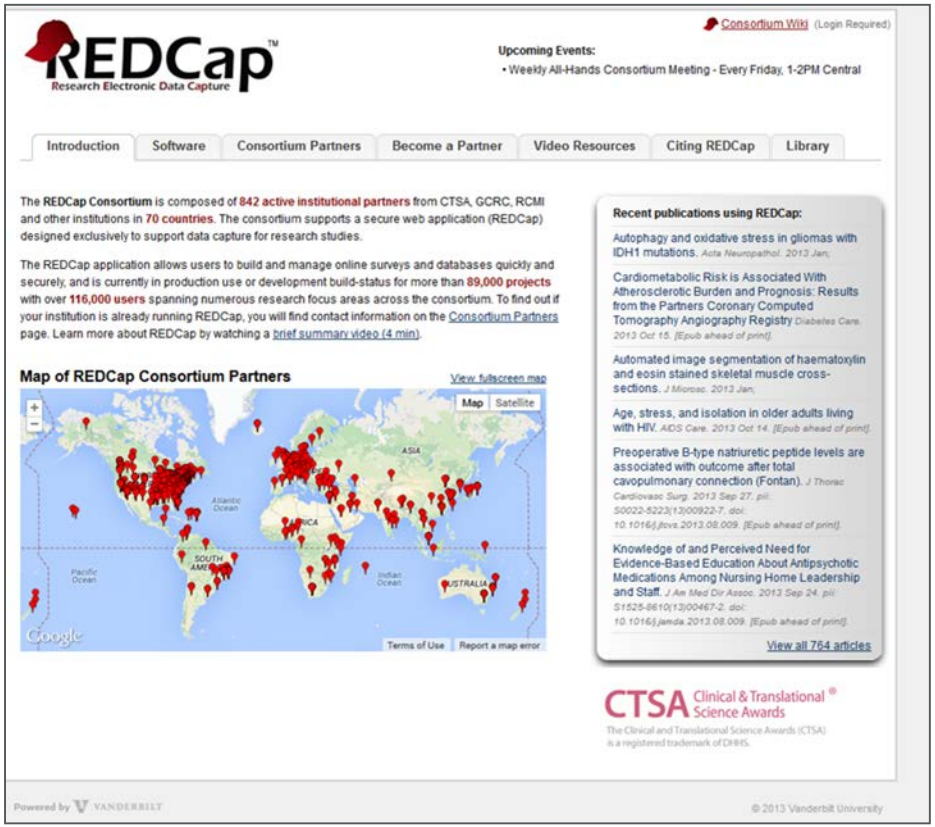


	A	B	C	D	E
1	casenum	year	month	age	race
2	1	78	1	4	black
3	2	78	12	1	black
4	3	78	3	0.8	black
5	4	78	8	54	black
6	5				
7	6				
8	7	78	12	0.5	black
9	8	78	2	0.1	black
10	9				
11	10	78	12	11	white
12	11	78	3	66	white
13	12	78	1	32	black
14	13	78	5	1	black

Tools for Reproducible Research: Data Capture

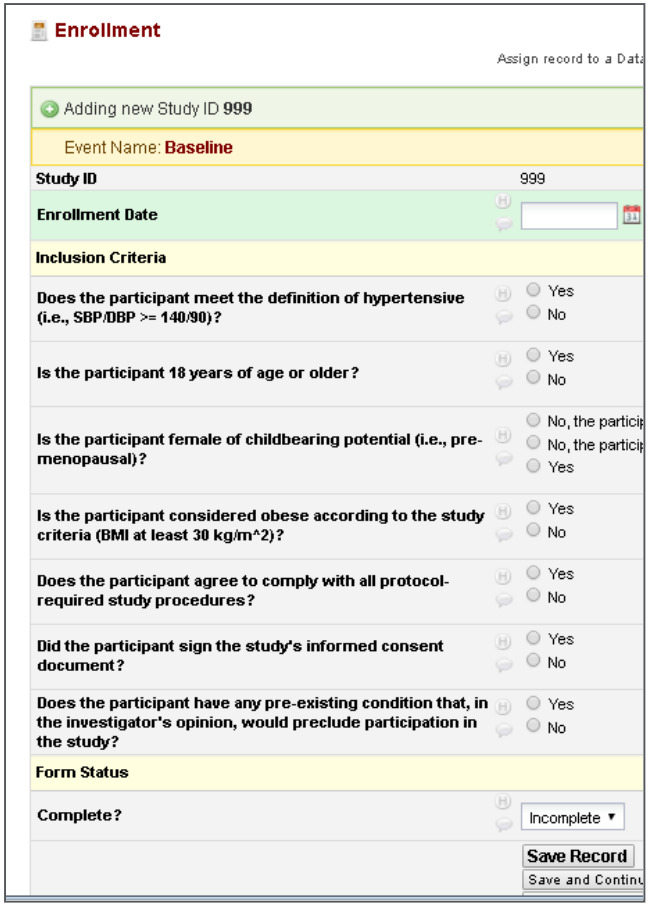
REDCap

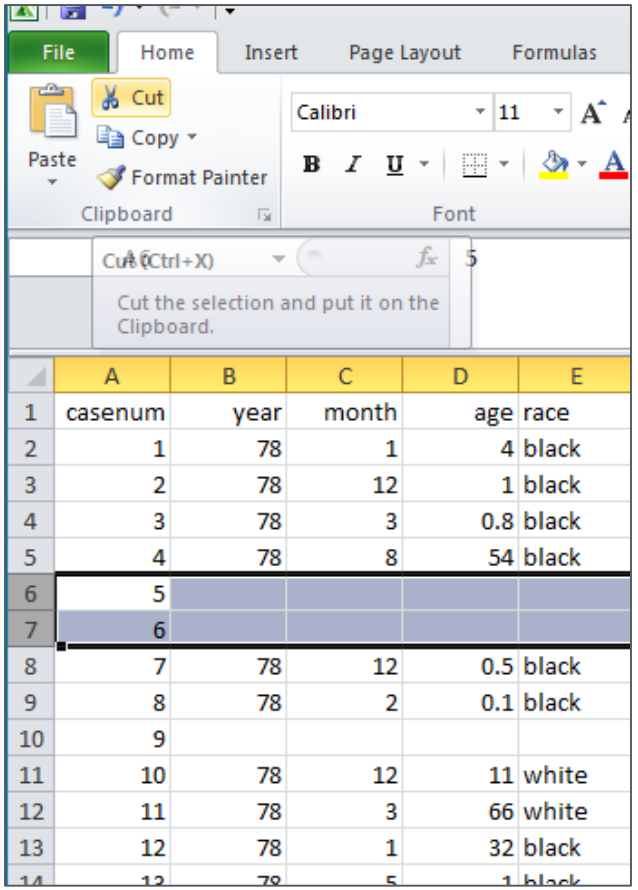
- **R**esearch **E**lectronic **D**ata **Cap**ture
- Secure web application
- http://project-redcap.org
- Features:
  - Rapid set-up
  - Web-based data collection
  - Data validation
  - Export to statistical programs
  - Supports HIPAA compliance



Tools for Reproducible Research: Data Capture

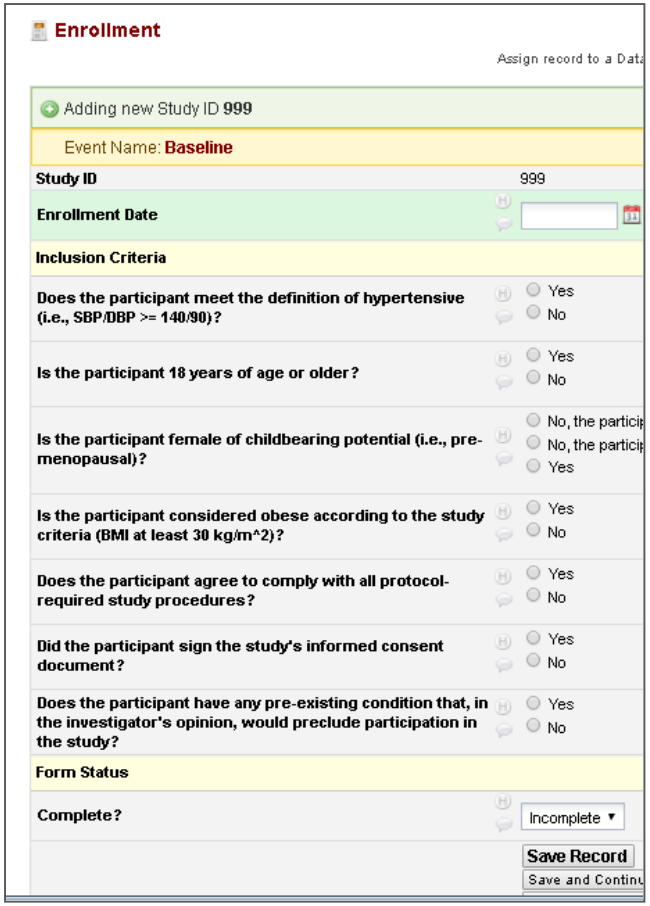
REDCap vs Excel

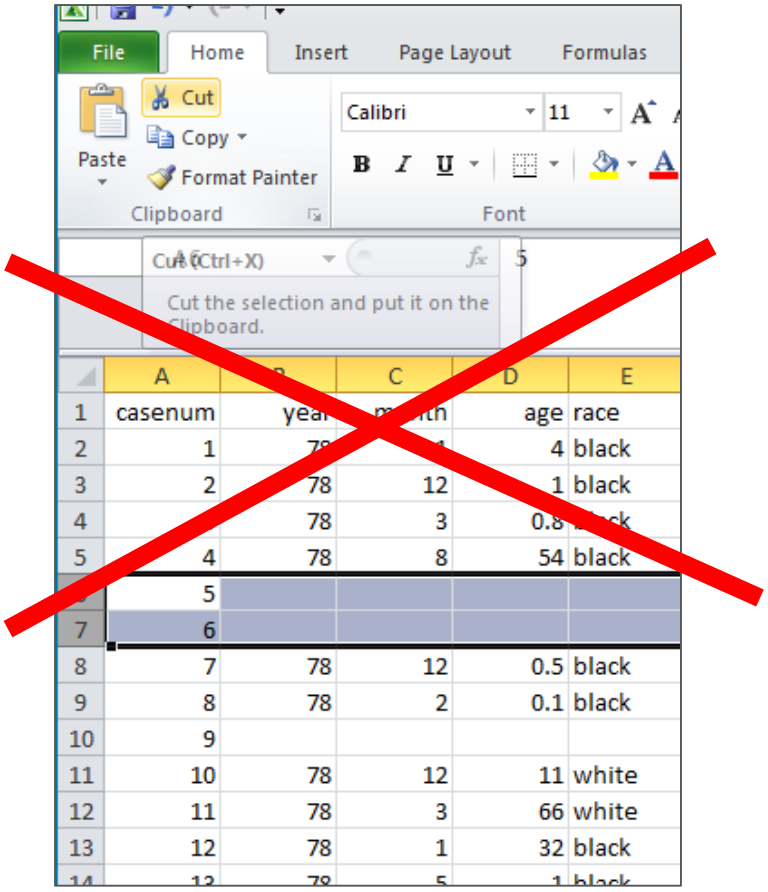




Tools for Reproducible Research: Data Capture

REDCap vs Excel





Tools for Reproducible Research: An Analogy

Staplers

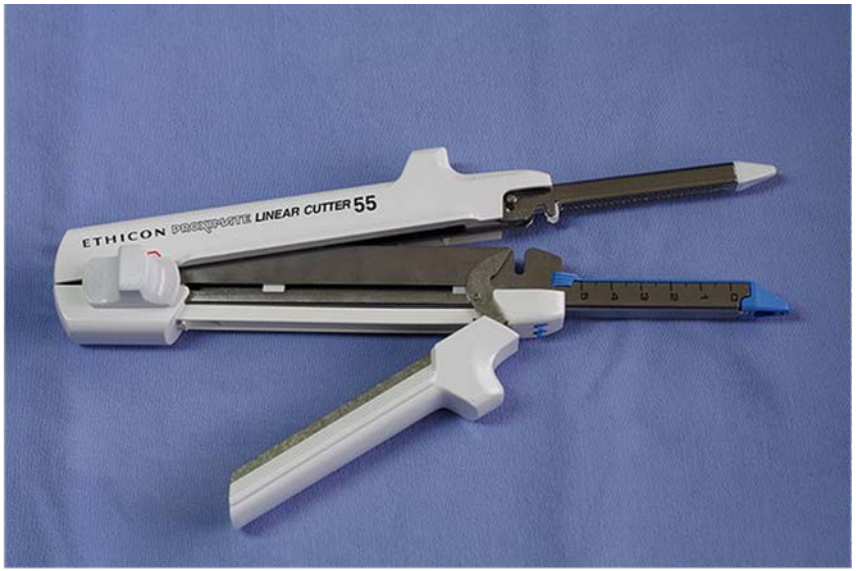
An all purpose office stapler:

Tools for Reproducible Research: An Analogy

Staplers

An all purpose office stapler:

A stapler designed for surgical procedures:



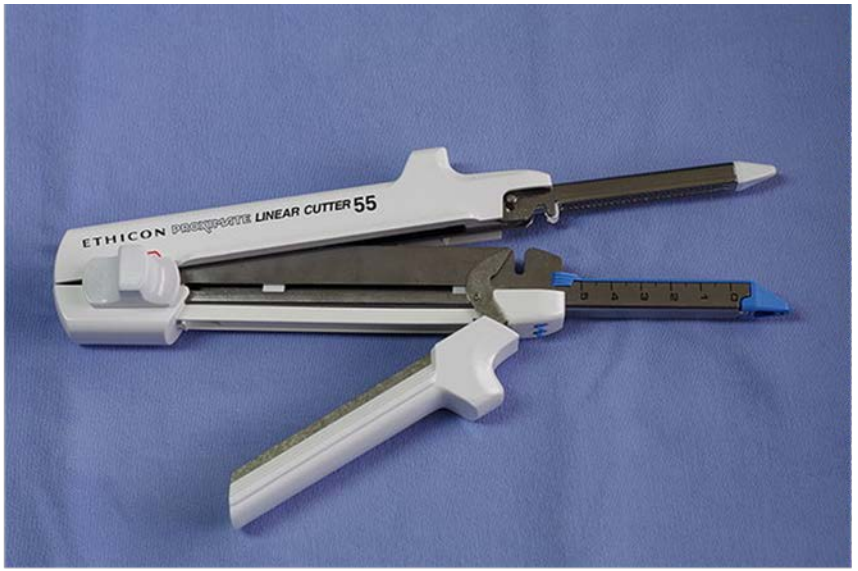
https://upload.wikimedia.org/wikipedia/commons/thumb/4/4a/Surgical_stapler_%26_cutter_linear.JPG/640px-Surgical_stapler_%26_cutter_linear.JPG

Tools for Reproducible Research: An Analogy

Staplers

An all purpose office stapler:

A stapler designed for surgical procedures:



https://upload.wikimedia.org/wikipedia/commons/thumb/4/4a/Surgical_stapler_%26_cutter_linear.JPG/640px-Surgical_stapler_%26_cutter_linear.JPG

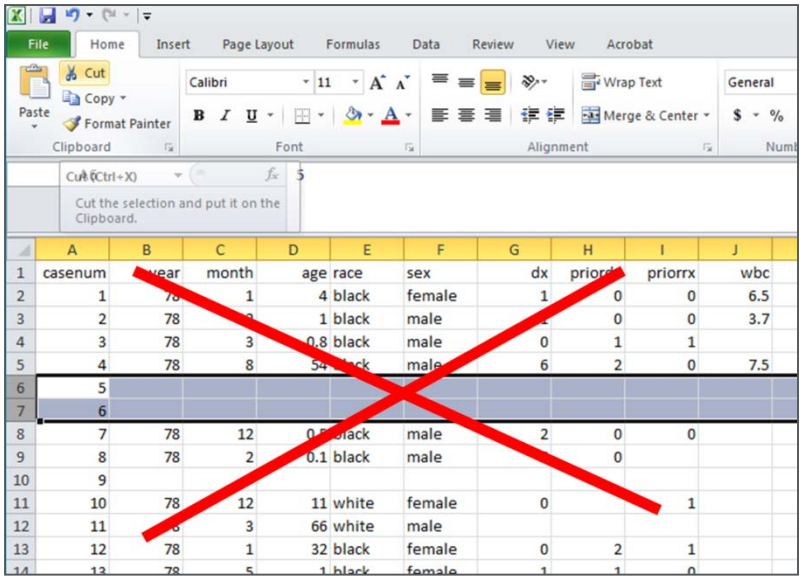
Would you ever use an office stapler to patch up a human being?

Tools for Reproducible Research: An Analogy

The Failures of Excel for Data Capture

You would never intentionally use an all purpose office stapler to patch up a human being.

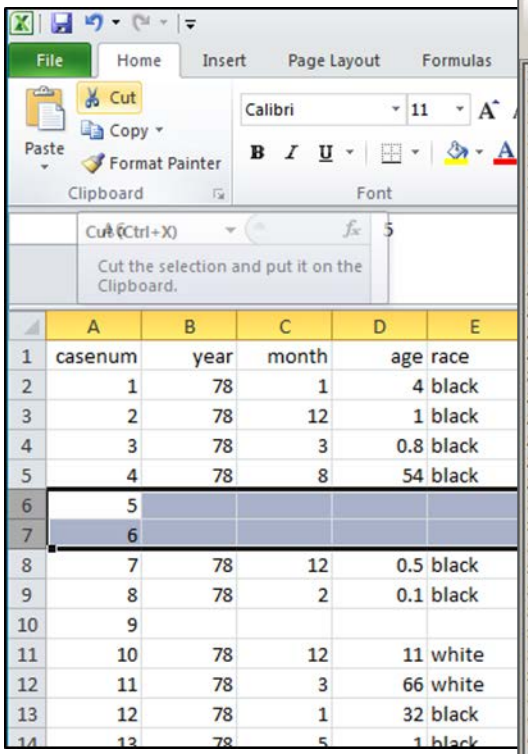
Why would you ever use an all purpose office spreadsheet program to capture potentially sensitive research data?



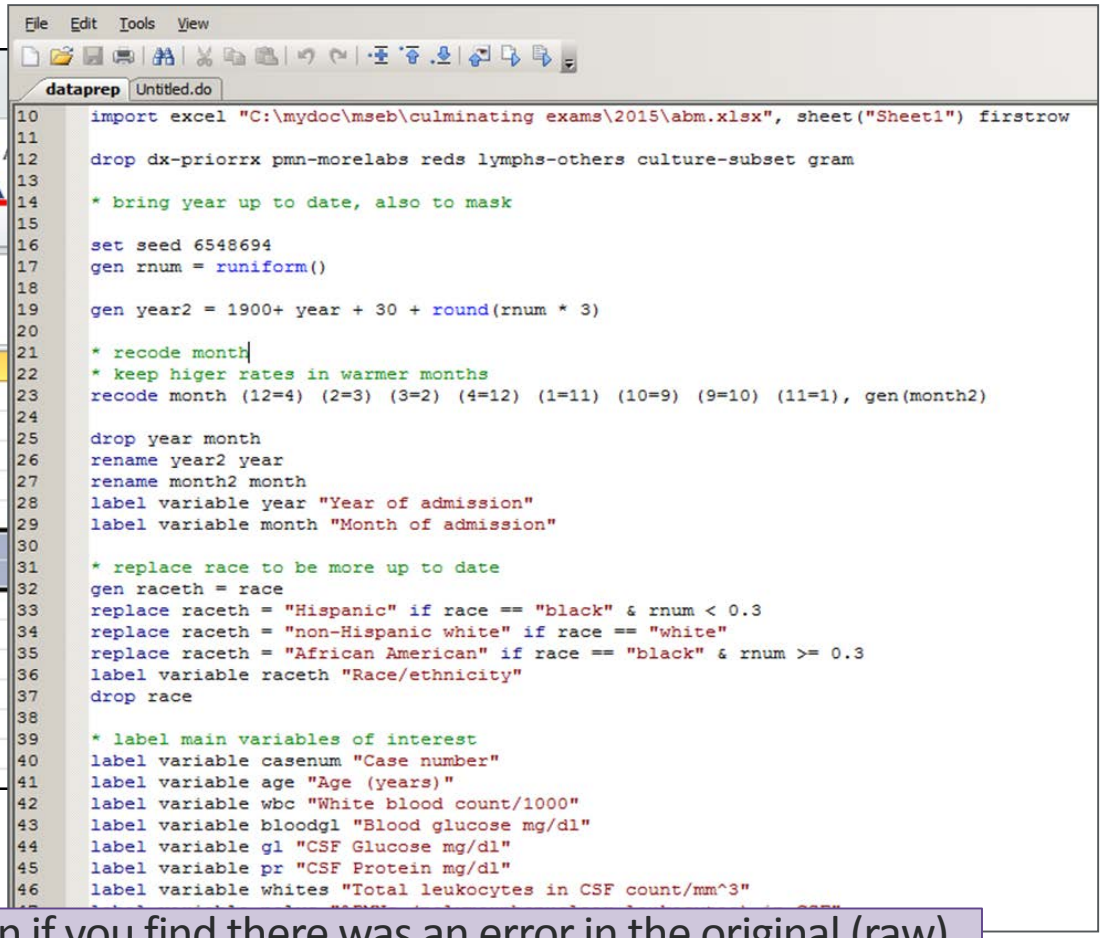
	A	B	C	D	E	F	G	H	I	J
1	casenum	year	month	age	race	sex	dx	priordx	priorrx	wbc
2	1	78	1	4	black	female	1	0	0	6.5
3	2	78	1	1	black	male	2	0	0	3.7
4	3	78	3	0.8	black	male	0	1	1	
5	4	78	8	54	black	male	6	2	0	7.5
6	5									
7	6									
8	7	78	12	0.5	black	male	2	0	0	
9	8	78	2	0.1	black	male		0		
10	9									
11	10	78	12	11	white	female	0		1	
12	11	78	3	66	white	male				
13	12	78	1	32	black	female	0	2	1	
14	13	78	5	1	black	female	1	1	0	

Tools for Reproducible Research: Data Preparation

Avoid cleaning and prepping your data in Excel



	A	B	C	D	E
1	casenum	year	month	age	race
2	1	78	1	4	black
3	2	78	12	1	black
4	3	78	3	0.8	black
5	4	78	8	54	black
6	5				
7	6				
8	7	78	12	0.5	black
9	8	78	2	0.1	black
10	9				
11	10	78	12	11	white
12	11	78	3	66	white
13	12	78	1	32	black
14	13	78	5	1	black



What will happen if you find there was an error in the original (raw) data? Will you know all the data manipulation steps to repeat (now or much later)?

Tools for Reproducible Research: Data Analysis

It's how we use the software

1. Point-and-click
2. Command line
3. Batch file (or text file of statistical code)

Tools for Reproducible Research: Data Analysis

It's how we use the software

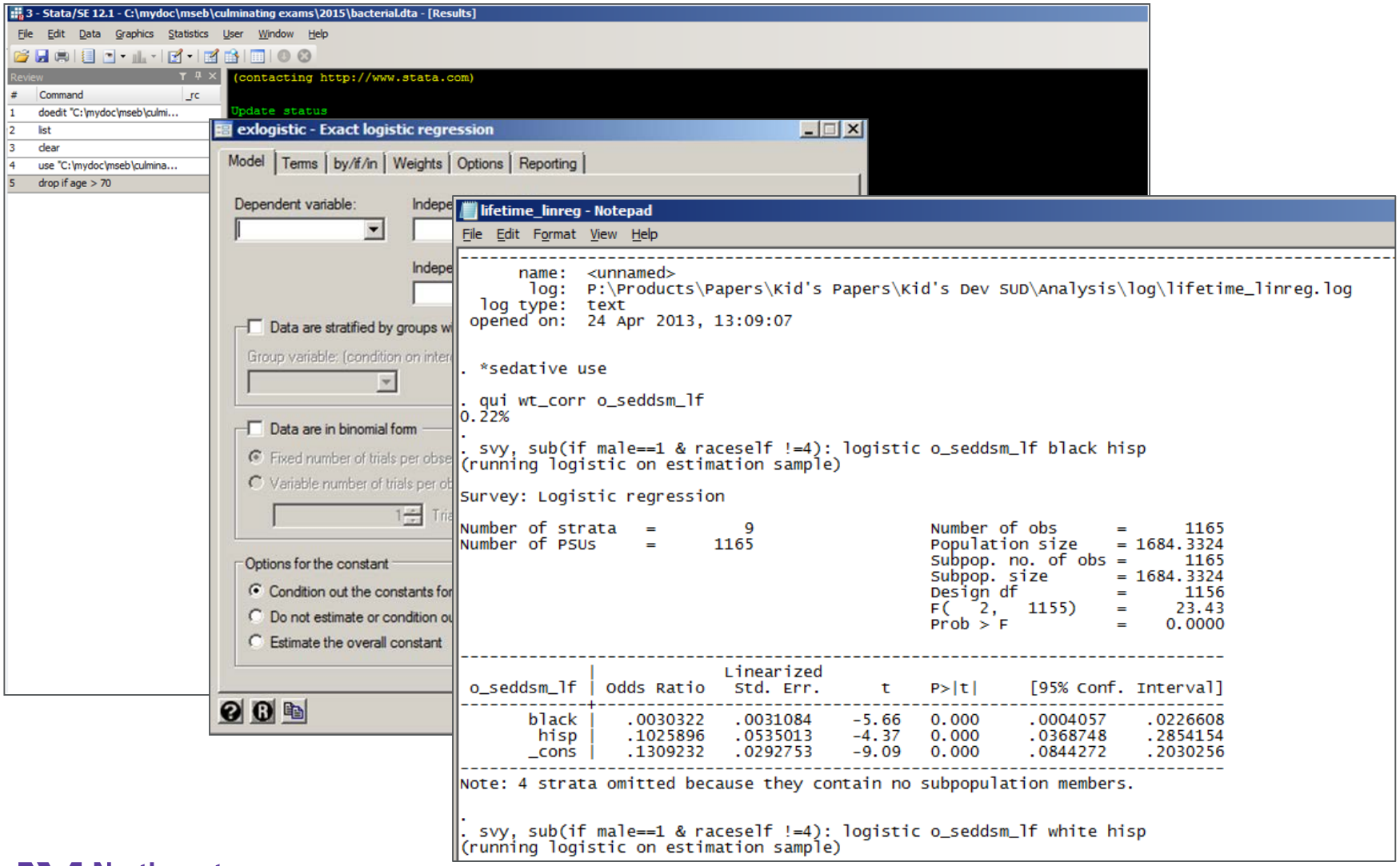
1. Point-and-click
2. Command line
3. Batch file (or text file of statistical code)

No matter how we use the software, we should keep a record of any and all manipulations to the data. Text files of written commands are preferable.

If we have to correct an error in the data, it can be documented in the code. All touches of the data should exist as a set of programming commands, or at the very least a copy of the execution of commands (e.g. "log" files in Stata).

Tools for Reproducible Research: Data Analysis

Stata: point-and-click is not completely incompatible with reproducibility

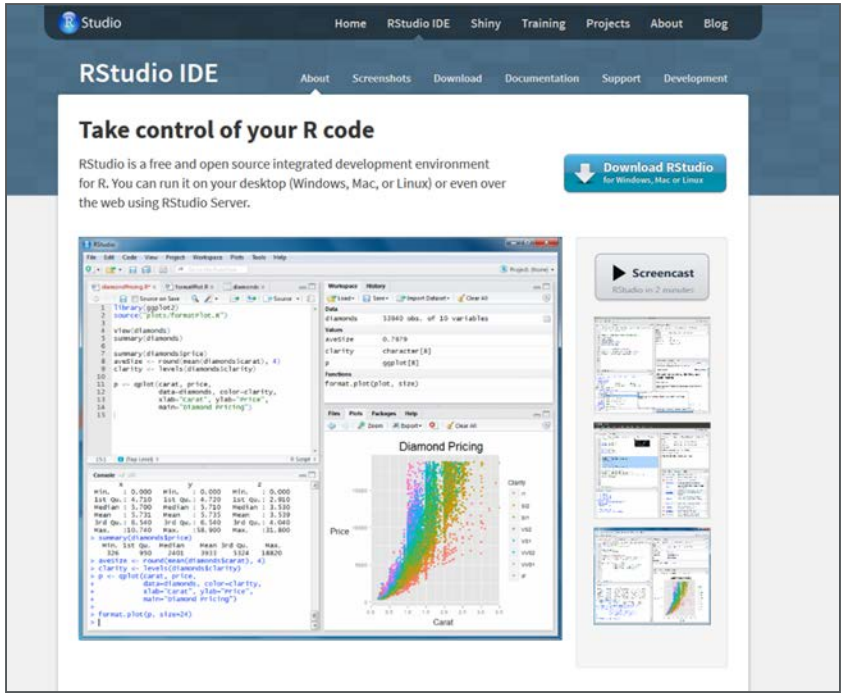


Northwestern Medicine®

Tools for Reproducible Research: Data Analysis

R and RStudio – not just for statisticians anymore

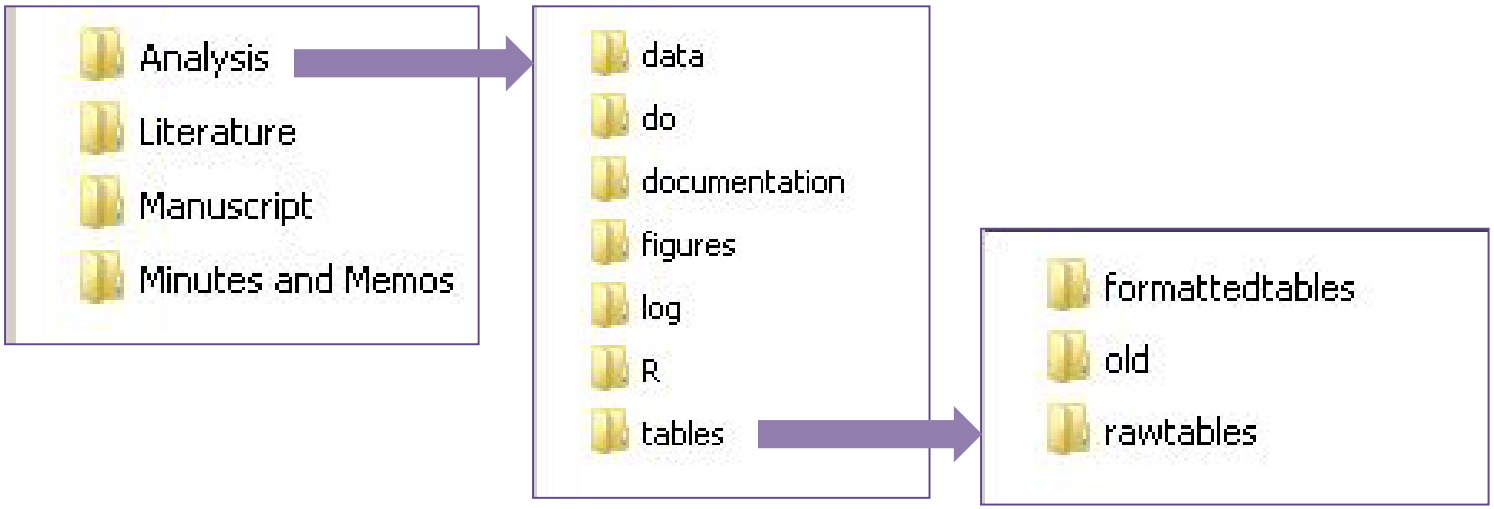
- R is open-source and free
- One of the main statistical programs
- Especially popular for ‘omics data
- Bundles together data and programs in packages for review and replication
- http://cran.us.r-project.org/
- RStudio
  - User friendly for non-statisticians and statisticians alike



Tools for Reproducible Research: Data Analysis

Organizing data & code

- Used REDCap to collect data
- Cleaned and analyzed data in SAS/R/Stata
- It will do you no good if we can't find or use our files
  - Which script do I run first?
  - Where did I store the data?
- Here's a system I use with my students and some collaborators
- But there is opportunity for a lot of improvement here

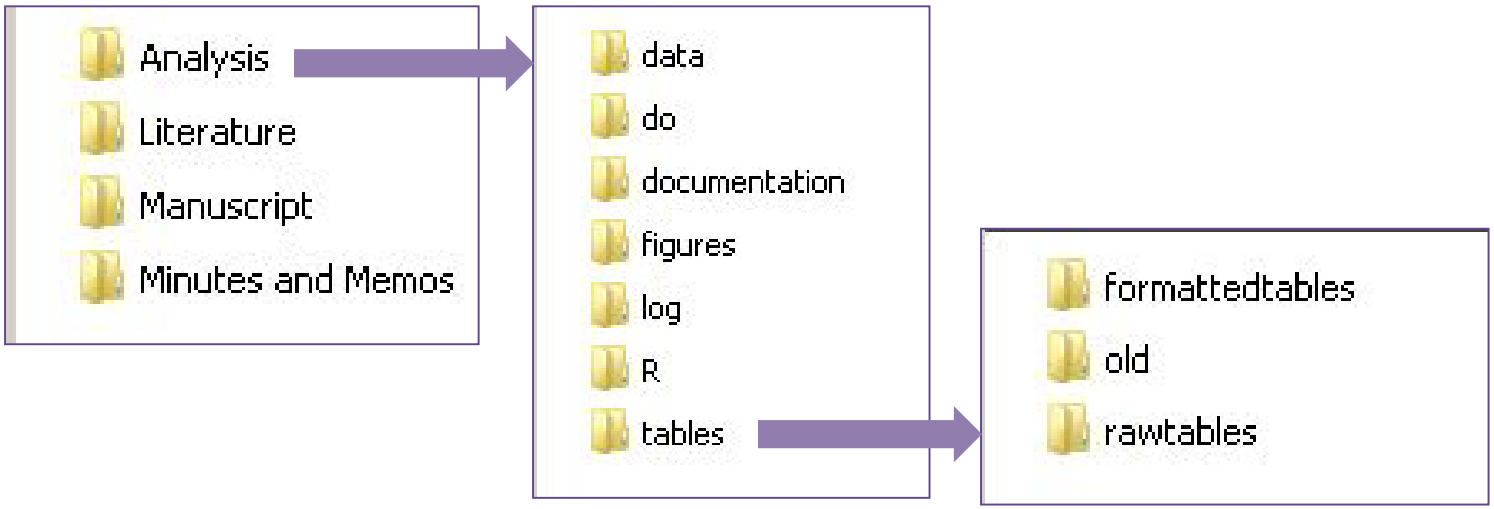


Tools for Reproducible Research: Data Analysis

Organizing data & code

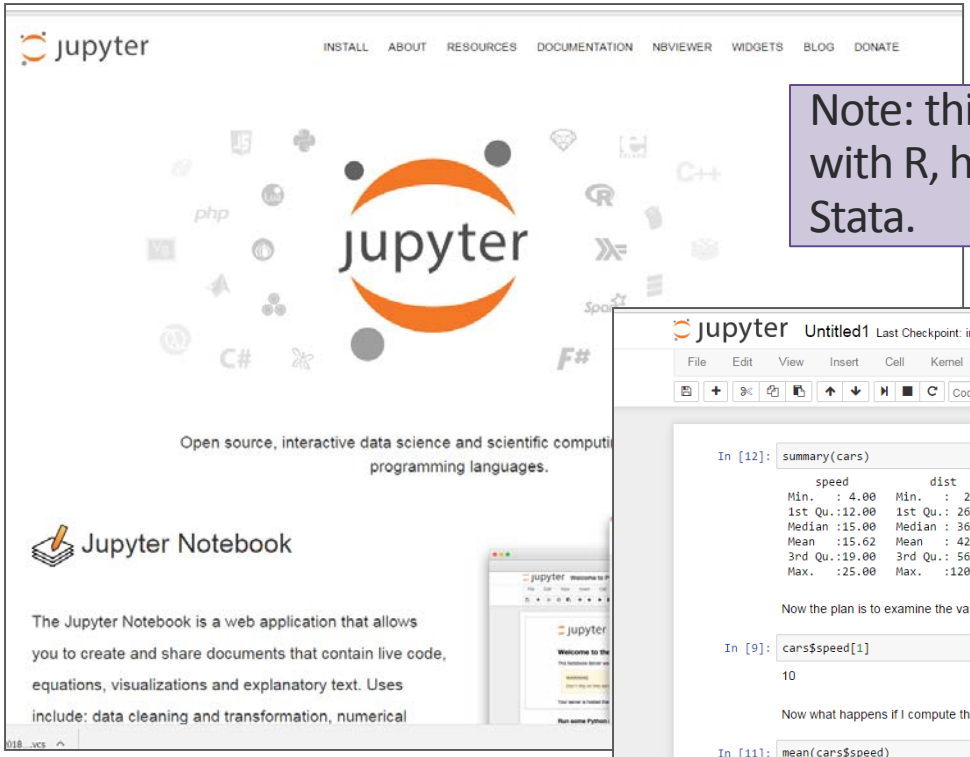
- Used REDCap to collect data
- Cleaned and analyzed data in SAS/R/Stata
- It will do you no good if we can't find or use our files
  - Which script do I run first?
  - Where did I store the data?
- Here's a system I use with my students and some collaborators
- But there is opportunity for a lot of improvement here

Need a system for version control! Can vary in sophistication from date in file name (simple) to automatic versioning (github, bitbucket, cvs).

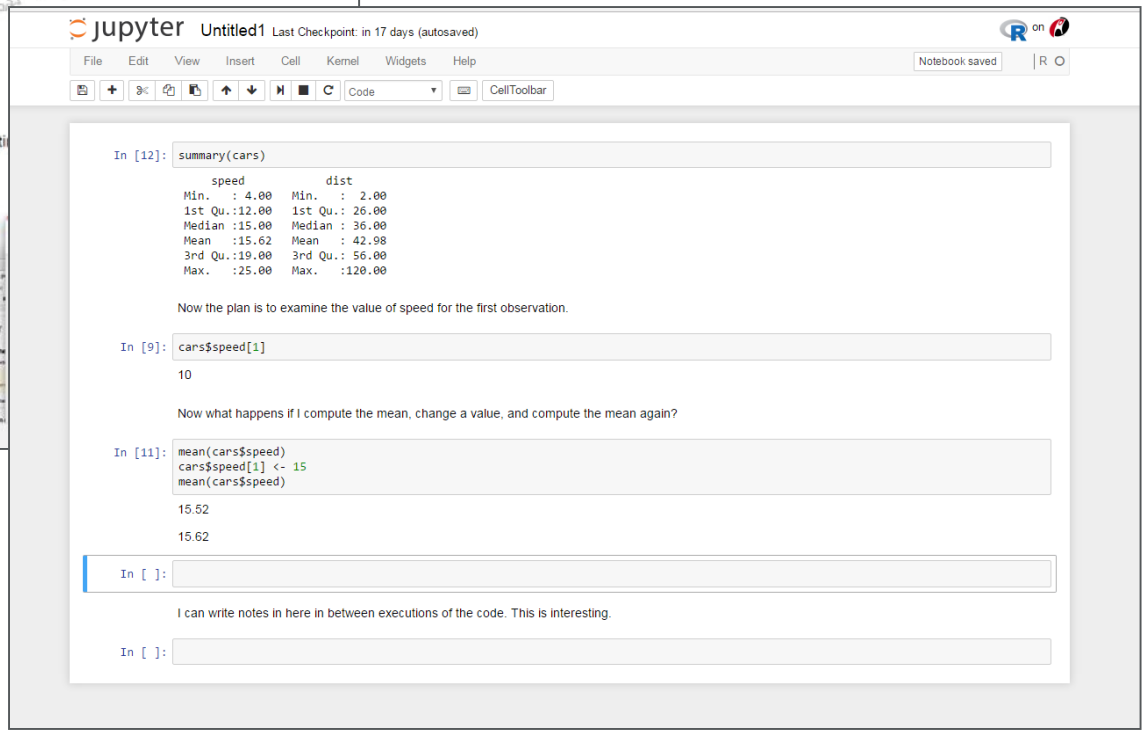


Tools for Reproducible Research: Data Analysis

jupyter for organizing data, code, communications and more



Note: this is for the fairly advanced user. Works with R, have seen ways to make it work with Stata.



Northwestern Medicine®

Tools for Reproducible Research: Data Analysis

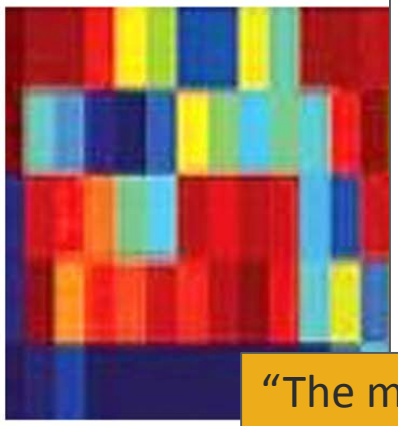
A disastrous story in why not to use Excel

Misconduct in science

An array of errors

Investigations into a case of alleged scientif
holes in the oversight of science and scienti

Sep 10th 2011 | From the print edition



ANIL POTTI, Joseph Nevins a
Carolina, garnered widesprea
of Medicine that they could pr
expression arrays, which log
tissue as a colourful picture (
that they had developed a sim
cultures of cancer cells, known as cell lines, to p
effective for an individual patient suffering from lu

At the time, this work looked like a tremendous a
idea that understanding the molecular specifics
treatment. The papers drew adulation from other
newspapers, including this one (see article), wro
organise a set of clinical trials of personalised treatments for lung and breast cancer.
Unbeknown to most people in the field, however, within a few weeks of the publication of the

The Annals of Applied Statistics
2009, Vol. 3, No. 4, 1309–1334
DOI: 10.1214/09-AOAS291

DERIVING CHEMOSENSITIVITY FROM CELL LINES: FORENSIC BIOINFORMATICS AND REPRODUCIBLE RESEARCH IN HIGH-THROUGHPUT BIOLOGY

BY KEITH A. BAGGERLY[1] AND KEVIN R. COOMBES[2]

University of Texas

High-throughput biological assays such as microarrays let us ask very detailed questions about how diseases operate, and promise to let us person-

show in five case studies that the results incorporate several simple errors that may be putting patients at risk. One theme that emerges is that the most common errors are simple (e.g., row or column offsets); conversely, it is our experience that the most simple errors are common. We then discuss steps we are taking to avoid such errors in our own investigations.

"The most simple problems are common." When using Excel, it is especially easy to make off-by-one errors (e.g. accidentally deleting a cell in one column), or mixing up group labels (e.g. swapping sensitive/resistant).

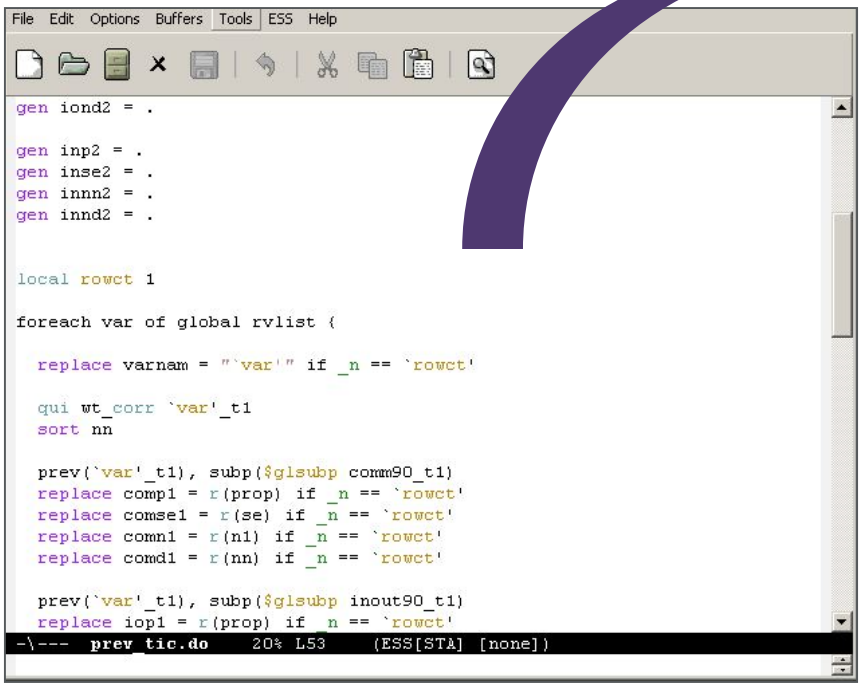
Tools for Reproducible Research: Manuscript Prep

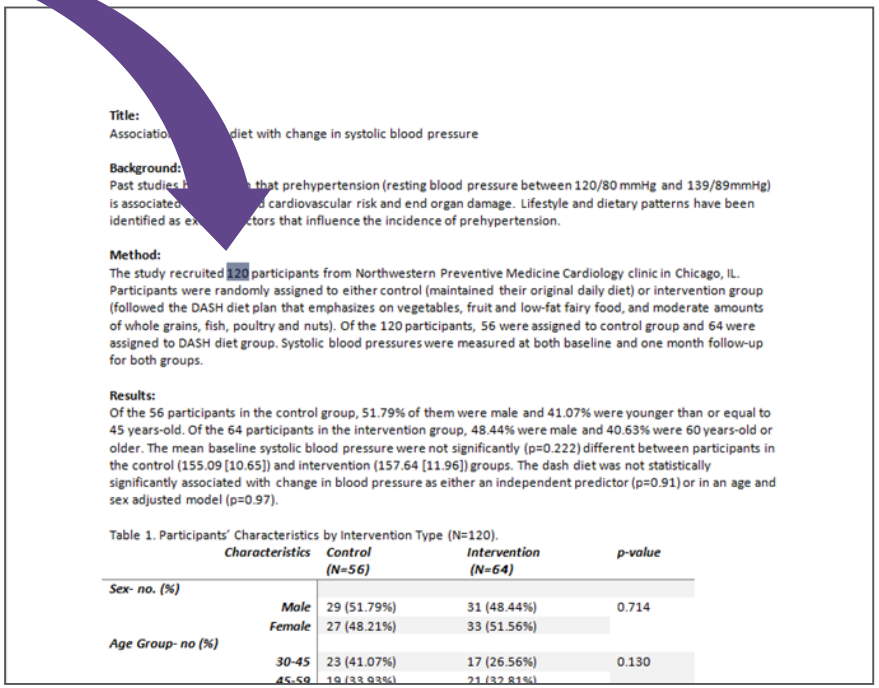
Publication: Connecting it all together

- Data Capture ✓
- Data Preparation ✓
- Data Analysis ✓
- Writing the manuscript – how do we link the estimates and our interpretation back to the code, variables, and data?

Tools for Reproducible Research: Manuscript Prep

Dynamic Documents





Title:

Associatio diet with change in systolic blood pressure

Background:

Past studies that prehypertension (resting blood pressure between 120/80 mmHg and 139/89mmHg) is associated d cardiovascular risk and end organ damage. Lifestyle and dietary patterns have been identified as ex tors that influence the incidence of prehypertension.

Method:

The study recruited 120 participants from Northwestern Preventive Medicine Cardiology clinic in Chicago, IL. Participants were randomly assigned to either control (maintained their original daily diet) or intervention group (followed the DASH diet plan that emphasizes on vegetables, fruit and low-fat fairy food, and moderate amounts of whole grains, fish, poultry and nuts). Of the 120 participants, 56 were assigned to control group and 64 were assigned to DASH diet group. Systolic blood pressures were measured at both baseline and one month follow-up for both groups.

Results:

Of the 56 participants in the control group, 51.79% of them were male and 41.07% were younger than or equal to 45 years-old. Of the 64 participants in the intervention group, 48.44% were male and 40.63% were 60 years-old or older. The mean baseline systolic blood pressure were not significantly (p=0.222) different between participants in the control (155.09 [10.65]) and intervention (157.64 [11.96]) groups. The dash diet was not statistically significantly associated with change in blood pressure as either an independent predictor (p=0.91) or in an age and sex adjusted model (p=0.97).

Table 1. Participants' Characteristics by Intervention Type (N=120).

Characteristics	*Control (N=56)*	*Intervention (N=64)*	*p-value*
Sex- no. (%)			
Male	29 (51.79%)	31 (48.44%)	0.714
Female	27 (48.21%)	33 (51.56%)	
Age Group- no (%)			
30-45	23 (41.07%)	17 (26.56%)	0.130
45-59	19 (33.93%)	21 (32.81%)	

- Rather than results being hard coded in a manuscript, they can be updated automatically when data or models change. For example, rather than re-entering updated odds ratios into a manuscript or table, everything is updated automatically, either when the document is opened or compiled.

Tools for Reproducible Research: Manuscript Prep

Dynamic Documents in SAS (ODS)

Merge and Update with Hopkins.sas *

```
COVER PAGE
*****************************************************************

ods escapechar="^";
OPTIONS NODATE Label;
%let num=9;
ods rtf file="PATHWAY\FILENAME &Sysdate..rtf";

title;
footnote;

/* Create a data set containing the desired title text */
data Test;
   text="Report:^n
        ^3n Status Update Report^10n &sysdate^5n Northwestern
run;

/* Insert blank lines (used to move the title text to the cente
footnote1 j=c "Confidential";
ods rtf text="^15n";

/* Output the title text */
proc report data=Test nowd noheader style(report)={rules=none f
     style(column)={font_weight=bold font_size=12pt just=c} ls=
run;

*****************************************************************
TABLE OF CONTENTS
*****************************************************************

ods rtf startpage=now;
footnote;

title1 "Table of Contents";
data Test;
  text="INTRODUCTION.....
        SUMMARY OF FILES RECEIVED.....
        SUMMARY OF DATA CLEANING.....
        DUPLICATE OBSERVATIONS.....
        SNPS COMPARED.....
        DATA SET CONCORDANCE.....
";
```

Report:

Status Update Report

29MAY15

Northwestern University

Confidential

Northwestern Medicine®

Tools for Reproducible Research: Manuscript Prep

Dynamic Documents with R Markdown

R markdown example.Rmd

Knit Word

```
---
title: "R Markdown Example"
author: "Leah Welty"
date: "April 6, 2016"
output: word_document
---

```{r setup, include=FALSE}
knitr::opts_chunk$set(echo = TRUE)
```
You can use R Markdown from within RStudio. You
language to indicate *italics* or **bold**. You

For example, if I want to see a summary of the *c
produces this:

```{r cars}
summary(cars)
```

I can also embed results directly in the text. F

That's pretty nice, because if I change something
how I'm changing the data:

```{r newmean}
cars$speed[1]
cars$speed[1] <- 10
```
So now if I generate the mean speed, it is `r mea

You can also include plots, and make tables using

R Markdown will take your plain text file and at
PDF, or MS Word. Pretty cool ... except ...

What happens when you send the Word document to a
abandoning R Markdown, or some unlucky person has
```

R Markdown Example

Leah Welty

April 6, 2016

You can use R Markdown from within RStudio. You write in a simple text editor, using the (fairly simple) Markdown language to indicate *italics* or **bold**. You can embed 'chunks' of R code and output in the document.

For example, if I want to see a summary of the *cars* dataset that comes standard with R, I can insert R code that produces this:

```
summary(cars)
```

```
##      speed           dist
##  Min.   : 4.0   Min.   :  2.00
##  1st Qu.:12.0   1st Qu.: 26.00
##  Median :15.0   Median : 36.00
##  Mean   :15.4   Mean   : 42.98
##  3rd Qu.:19.0   3rd Qu.: 56.00
##  Max.   :25.0   Max.   :120.00
```

I can also embed results directly in the text. For example, the median speed is 15.4.

That's pretty nice, because if I change something about the data, then that number can be automatically updated. This is how I'm changing the data:

```
cars$speed[1]
```

```
## [1] 4
```

```
cars$speed[1] <- 10
```

So now if I generate the mean speed, it is 15.52.

You can also include plots, and make tables using R Markdown.

R Markdown will take your plain text file and at the touch of a button, insert all the R output then turn it in to HTML, PDF, or MS Word. Pretty cool ... except ...

What happens when you send the Word document to a collaborator, and they mark it up in track changes? [Hint: You end up abandoning R Markdown, or some unlucky person has to go back and insert all those changes in Markdown]

Northwestern Medicine®

Tools for Reproducible Research: Manuscript Prep

Dynamic Documents with R Markdown

R markdown example.Rmd

Knit Word

```
---
title: "R Markdown Example"
author: "Leah Welty"
date: "April 6, 2016"
output: word_document
---

```{r setup, include=FALSE}
knitr::opts_chunk$set(echo = TRUE)
```
You can use R Markdown from within RStudio. You
language to indicate *italics* or **bold**. You

For example, if I want to see a summary of the *c
produces this:

```{r cars}
summary(cars)
```

I can also embed results directly in the text. F

That's pretty nice, because if I change something
how I'm changing the data:

```{r newmean}
cars$speed[1]
cars$speed[1] <- 10
```
So now if I generate the mean speed, it is `r mea

You can also include plots, and make tables using

R Markdown will take your plain text file and at
PDF, or MS Word. Pretty cool ... except ...

What happens when you send the Word document to a
abandoning R Markdown, or some unlucky person has
```

R Markdown Example

Leah Welty

April 6, 2016

You can use R Markdown from within RStudio. You write in a simple text editor, using the (fairly simple) Markdown language to indicate *italics* or **bold**. You can embed 'chunks' of R code and output in the document.

For example, if I want to see a summary of the *cars* dataset that comes standard with R, I can insert R code that produces this:

```
summary(cars)
```

```
##      speed           dist
##  Min.   : 4.0   Min.   :  2.00
##  1st Qu.:12.0   1st Qu.: 26.00
##  Median :15.0   Median : 36.00
##  Mean   :15.4   Mean   : 42.98
##  3rd Qu.:19.0   3rd Qu.: 56.00
##  Max.   :25.0   Max.   :120.00
```

I can also embed results directly in the text. For example, the median speed is 15.4.

That's pretty nice, because if I change something about the data, then that number can be automatically updated. This is how I'm changing the data:

```
cars$speed[1]
```

```
## [1] 4
```

```
cars$speed[1] <- 10
```

So now if I generate the mean speed, it is 15.52.

You can also include plots, and make tables using R Markdown.

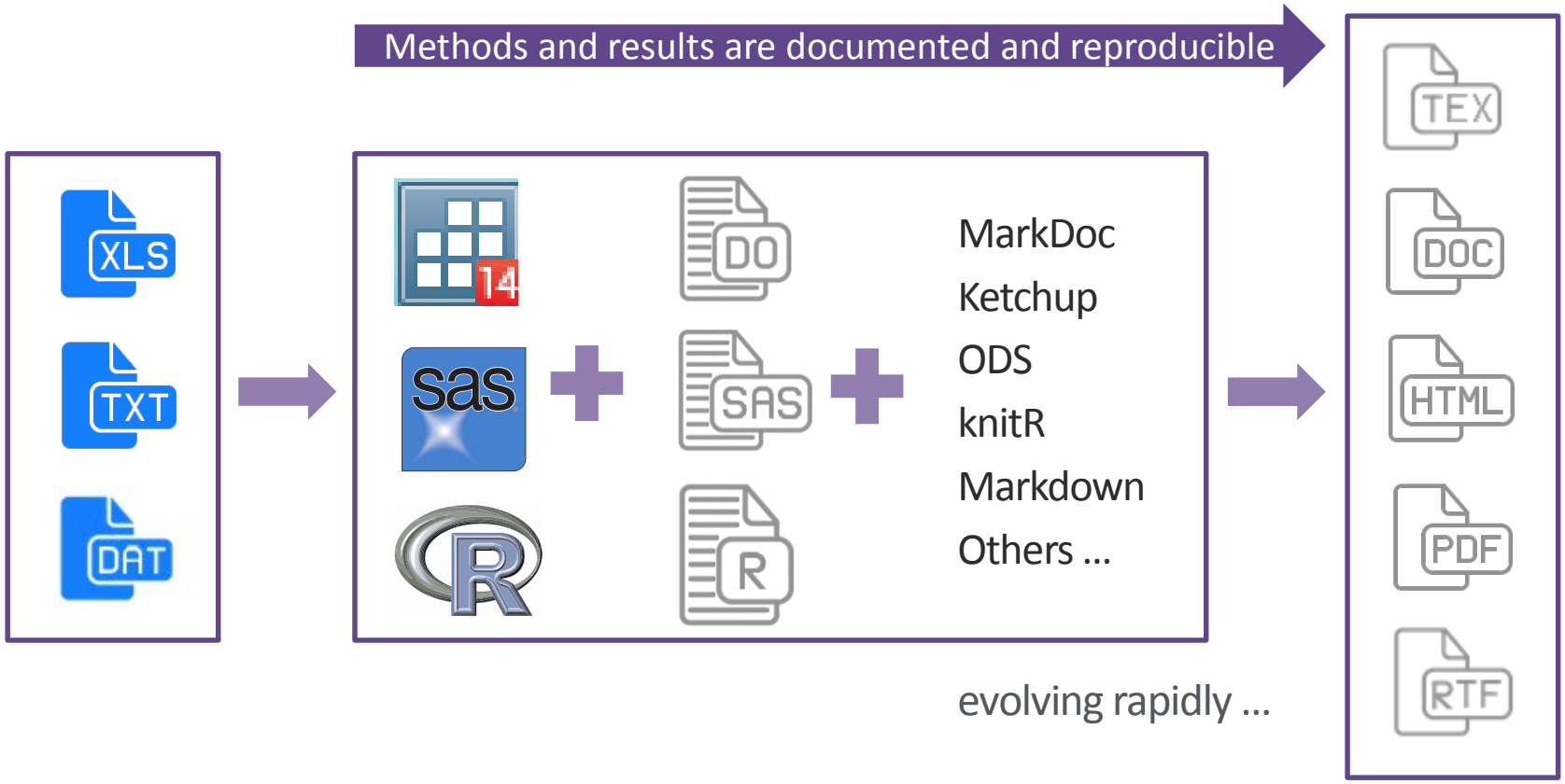
R Markdown will take your plain text file and at the touch of a button, insert all the R output then turn it in to HTML, PDF, or MS Word. Pretty cool ... except ...

What happens when you send the Word document to a collaborator, and they mark it up in track changes? [Hint: You end up abandoning R Markdown, or some unlucky person has to go back and insert all those changes in Markdown]

Northwestern Medicine®

Tools for Reproducible Research: Manuscript Prep

Existing tools for Dynamic Documents



Tools for Reproducible Research: Manuscript Prep

The Problem with Dynamic Documents: Text Files and Collaborators

Current tools require writing within a text editor. For example, a Markdown document looks something like this:

```
author: "Leah Welty"
date: "July 27, 2006"
output: word_document
---

```{r setup, include=FALSE}
knitr::opts_chunk$set(echo = TRUE)
```

You can use R Markdown from within RStudio. You write in a simple text editor, using the (fairly simple) Markdown language to indicate *italics* output in the document.

For example, if I want to see a summary of the *cars* dataset that comes standard with R, I can insert R code that produces this:

```{r cars}
summary(cars)
```

I can also embed results directly in the text. For example, the median speed is `r mean(cars$speed)`.

That's pretty nice, because if I change something about the data, then that number can be automatically updated.

Another recent thing is that I can actually call and run Stata code from this interface. Neat, but I still have a problem ...
```

Tools for Reproducible Research: Manuscript Prep

The Problem with Dynamic Documents: Text Files and Collaborators

Current tools require writing within a text editor. For example, a Markdown document looks something like this:

```
author: "Leah Welty"
date: "July 27, 2006"
output: word_document
---

```{r setup, include=FALSE}
knitr::opts_chunk$set(echo = TRUE)
```

You can use R Markdown from within RStudio. You write in a simple text editor, using the (fairly simple) Markdown language to indicate *italics* output in the document.

For example, if I want to see a summary of the *cars* dataset that comes standard with R, I can insert R code that produces this:

```{r cars}
summary(cars)
```

I can also embed results directly in the text. For example, the median speed is `r mean(cars$speed)`.

That's pretty nice, because if I change something about the data, then that number can be automatically updated.

Another recent thing is that I can actually call and run Stata code from this interface. Neat, but I still have a problem ...
```

Are you or your non-technical collaborators willing to work this way? I'm happy to work in text editors, but my collaborators (primarily clinicians and social scientists) are not.

Tools for Reproducible Research: Manuscript Prep

A Problem for Dynamic Documents: Track Changes

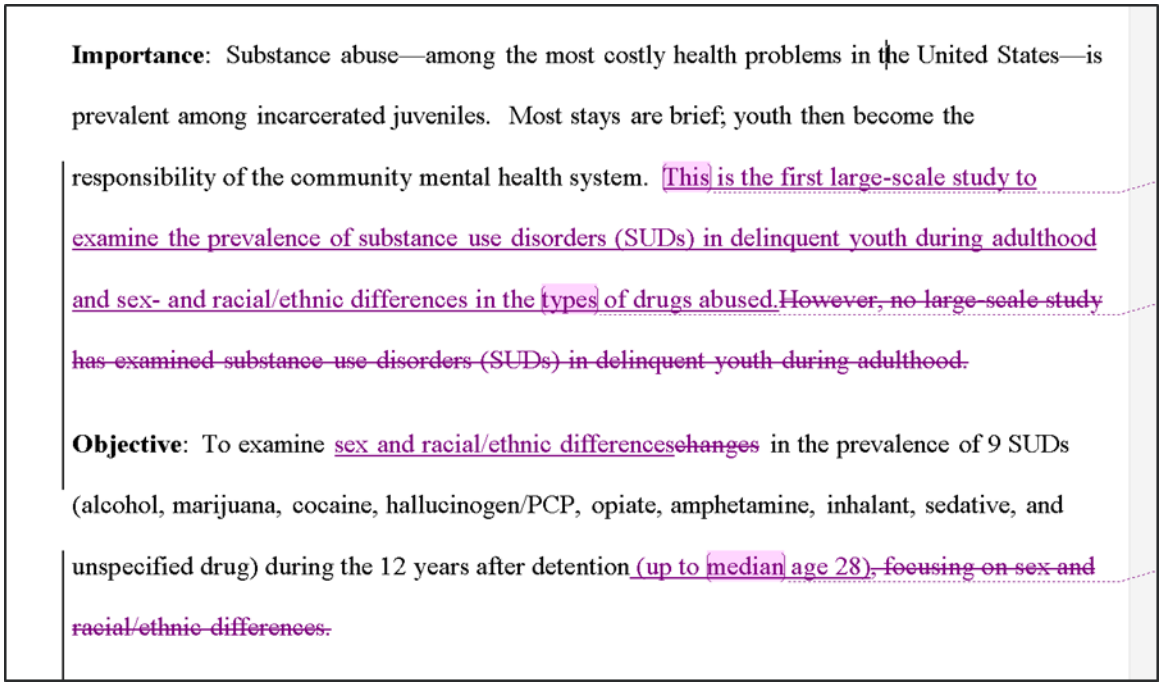
I create a dynamic document, generate the Word file and send it to collaborators.

Tools for Reproducible Research: Manuscript Prep

A Problem for Dynamic Documents: Track Changes

I create a dynamic document, generate the Word file and send it to collaborators.

They send back:



Importance: Substance abuse—among the most costly health problems in the United States—is prevalent among incarcerated juveniles. Most stays are brief; youth then become the responsibility of the community mental health system. This is the first large-scale study to examine the prevalence of substance use disorders (SUDs) in delinquent youth during adulthood and sex- and racial/ethnic differences in the types of drugs abused. ~~However, no large-scale study has examined substance use disorders (SUDs) in delinquent youth during adulthood.~~

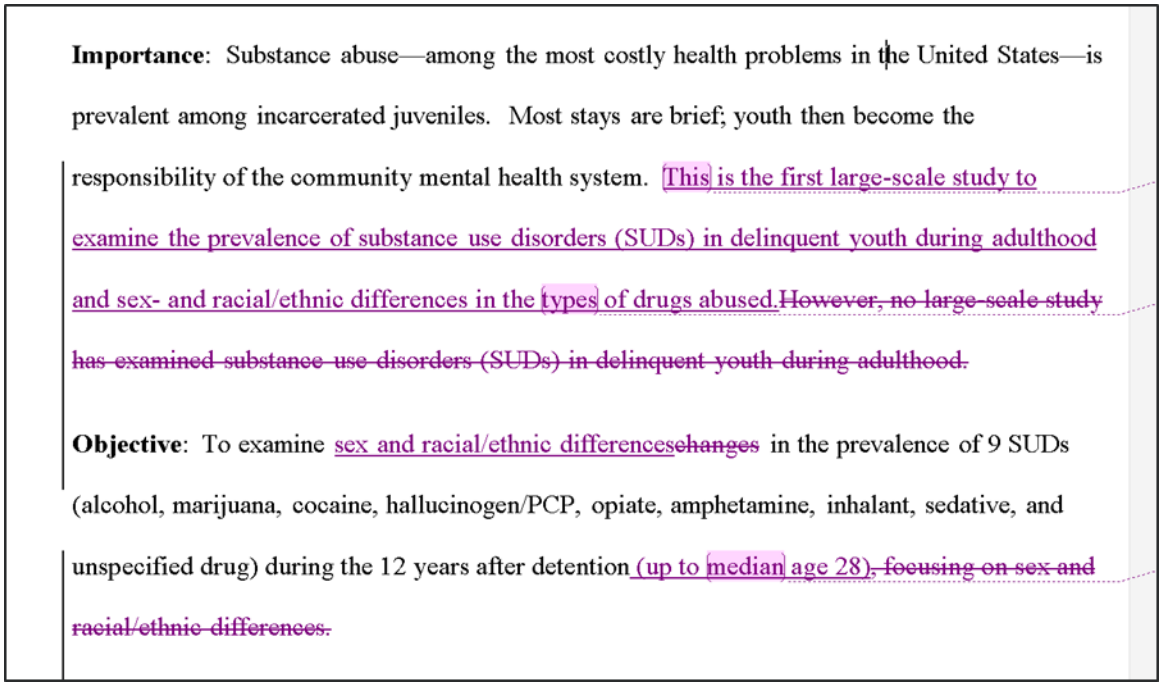
Objective: To examine sex and racial/ethnic differences~~changes~~ in the prevalence of 9 SUDs (alcohol, marijuana, cocaine, hallucinogen/PCP, opiate, amphetamine, inhalant, sedative, and unspecified drug) during the 12 years after detention (up to median age 28)~~, focusing on sex and racial/ethnic differences.~~

Tools for Reproducible Research: Manuscript Prep

A Problem for Dynamic Documents: Track Changes

I create a dynamic document, generate the Word file and send it to collaborators.

They send back:



Importance: Substance abuse—among the most costly health problems in the United States—is prevalent among incarcerated juveniles. Most stays are brief; youth then become the responsibility of the community mental health system. This is the first large-scale study to examine the prevalence of substance use disorders (SUDs) in delinquent youth during adulthood and sex- and racial/ethnic differences in the types of drugs abused. ~~However, no large-scale study has examined substance use disorders (SUDs) in delinquent youth during adulthood.~~

Objective: To examine sex and racial/ethnic differences~~changes~~ in the prevalence of 9 SUDs (alcohol, marijuana, cocaine, hallucinogen/PCP, opiate, amphetamine, inhalant, sedative, and unspecified drug) during the 12 years after detention (up to median age 28), ~~focusing on sex and racial/ethnic differences.~~

I have two (bad) choices:

1. Continue in Word, and loose the dynamic nature of the document.
2. Re-enter all of their changes in my source file.

Tools for Reproducible Research: Manuscript Prep

A Problem for Dynamic Documents: MS Word is Ubiquitous

The NEW ENGLAND JOURNAL of MEDICINE

*"All text…should be in one double-spaced electronic document (preferably a **Word Doc**)"*

JAMA The Journal of the American Medical Association

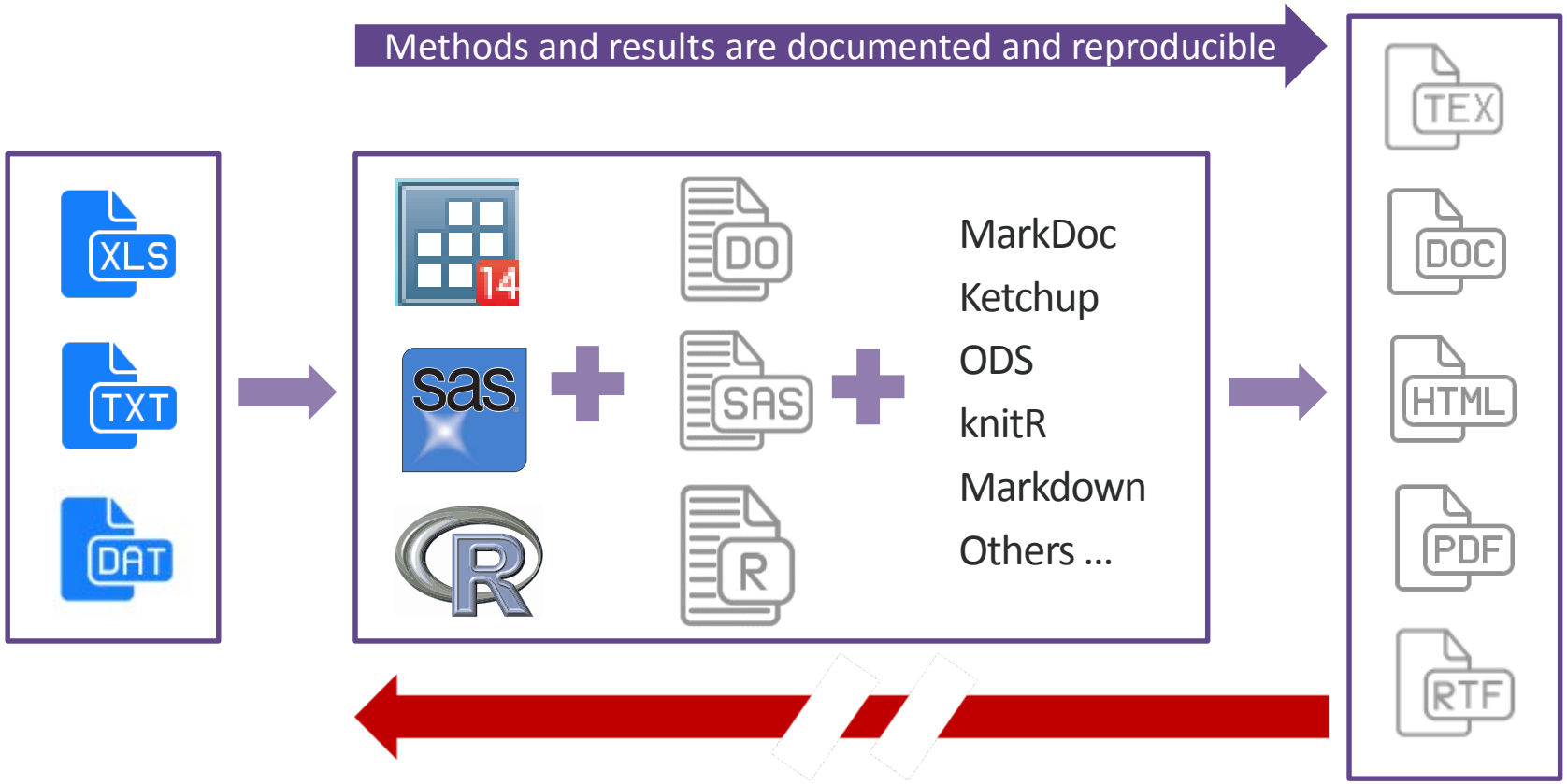
*"…acceptable manuscript file formats include **Word** and WordPerfect. Do not submit your manuscript in PDF format."*

BMJ British Medical Journal

*"The manuscript must be submitted as a **Word** document. PDF is not accepted."*

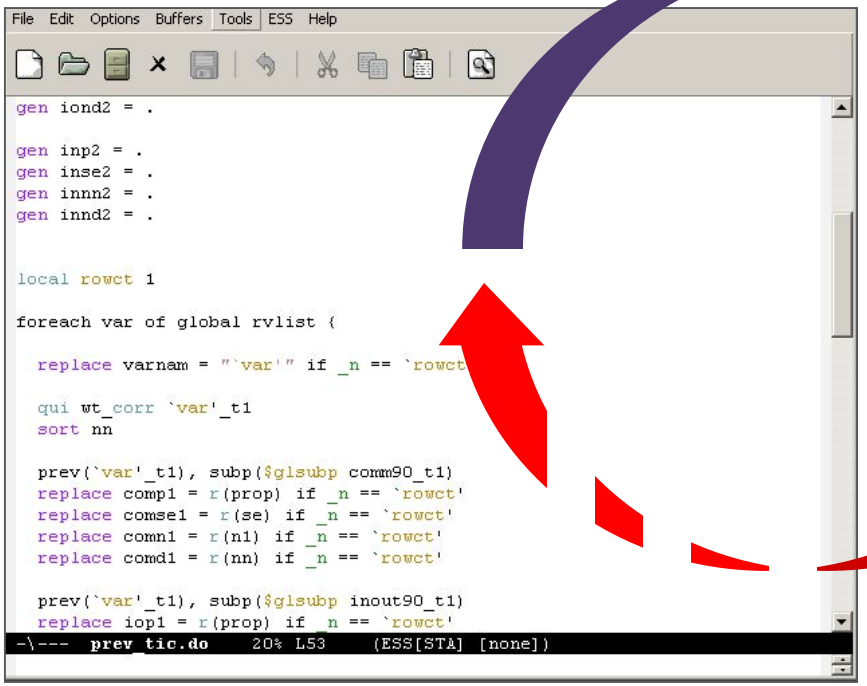
Tools for Reproducible Research: Manuscript Prep

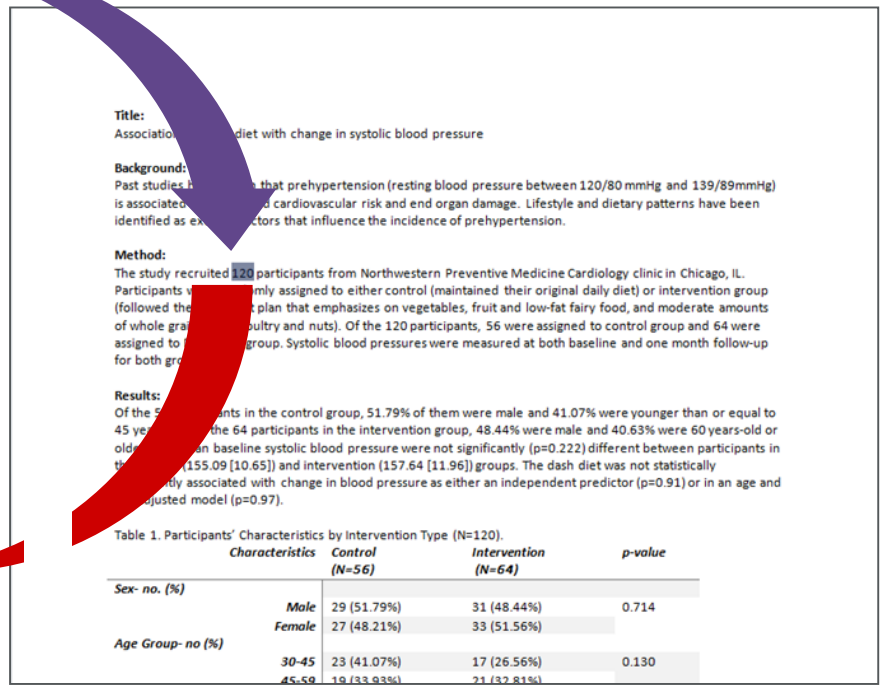
Limitations of Existing Tools for Dynamic Documents



Tools for Reproducible Research: Manuscript Prep

Limitations of Existing Tools for Dynamic Documents





- Rather than results being hard coded in a manuscript, they can be updated automatically when data or models change. For example, rather than re-entering updated odds ratios into a manuscript or table, everything is updated automatically, either when the document is opened or compiled.

Northwestern Medicine®

Tools for Reproducible Research: Manuscript Prep

StatTag for Dynamic Documents with Microsoft Word

Full disclosure:

StatTag was developed here at Northwestern by:

Luke V. Rasmussen, Lead Software Developer
Abigail S. Baldridge, Senior Statistical Analyst
Eric W. Whitley, Software Developer
Leah J. Welty, Biostatistician

Development of StatTag was supported, in part, by the National Institutes of Health's National Center for Advancing Translational Sciences, Grant Number UL1TR001422.

Tools for Reproducible Research: StatTag

What makes StatTag different than other programs?

- StatTag is a free plug-in for Microsoft Word
  - Connects Stata or SAS code and Word document
  - You and your collaborators can work from the same Word document without breaking links between the code and data
  - Can work separately on code and the Word document

- User-friendly, easy learning curve
  - StatTag menu consistent with Word layout
  - EndNote:Citations as StatTag:Results

StatTag: An Introduction

StatTag-Updates-720.wmv

Northwestern Medicine®

StatTag: An Introduction

StatTag-Tagging-720.wmv

Northwestern Medicine®

Getting to know StatTag

Can I use StatTag with multiple code files?

- Yes! StatTag can connect to multiple ".do" or ".sas" files.

- Future versions will connect to R. You can connect both Stata and SAS files to the same document.

Getting to know StatTag

How does StatTag work when I share the Word document with collaborators?

If I have…	I can…		
	Review/edit manuscript text	View code associated with a tag	Insert or update a tag
Microsoft Word	✓	✗	✗

Getting to know StatTag

How does StatTag work when I share the Word document with collaborators?

If I have…	I can…		
	Review/edit manuscript text	View code associated with a tag	Insert or update a tag
Microsoft Word	✓	✗	✗
+ StatTag and Stata/SAS code	✓	✓	✗

Getting to know StatTag

How does StatTag work when I share the Word document with collaborators?

If I have…	I can…		
	Review/edit manuscript text	View code associated with a tag	Insert or update a tag
Microsoft Word	✓	✗	✗
+ StatTag and Stata/SAS code	✓	✓	✗
+ Stata/SAS and Data	✓	✓	✓

Getting to know StatTag

Is it only for Word and Stata/SAS/R on Windows?

- No! **The first public releases are for Windows and Stata/SAS**.
- A Mac version of StatTag for Stata is coming soon.

	Stata	SAS	R
Windows	✓	✓	April 2017
Mac	Early 2017	X	June 2017

Getting to know StatTag

Can I see all the tags in a document?

StatTag

- Yes! Inserted tags are highlighted when they are clicked on.
- Future versions will include a "highlight all tags" function to quickly find any inserted tags in a document.

CONCLUSIONS:
Intervention X was not statistically significantly associated with a reduction in S
placebo control. Longer term follow up may be needed to assess if interventio
time.

Table 1. Participant Characteristics (N=120).

Characteristic, N (%)	Control (N=56)		Intervention (N=64)		P-value*
Male	29.00	0.24	31.00	0.26	0.71
Female	27.00	0.22	33.00	0.28	.
30-45 Years	23.00	0.19	17.00	0.14	0.13
45-59 Years	19.00	0.16	21.00	0.17	.
60+ Years	14.00	0.12	26.00	0.22	.
SBP Before**	155.09	10.65	157.64	11.96	0.22
SBP After**	149.80	13.78	152.72	14.48	0.26
SBP Change**	-5.29	15.51	-4.92	17.82	0.91

* Chi-squared or t-test
** Presented as mean (sd)

Northwestern Medicine®

Getting to know StatTag

What about data security (e.g. PII, PHI)?

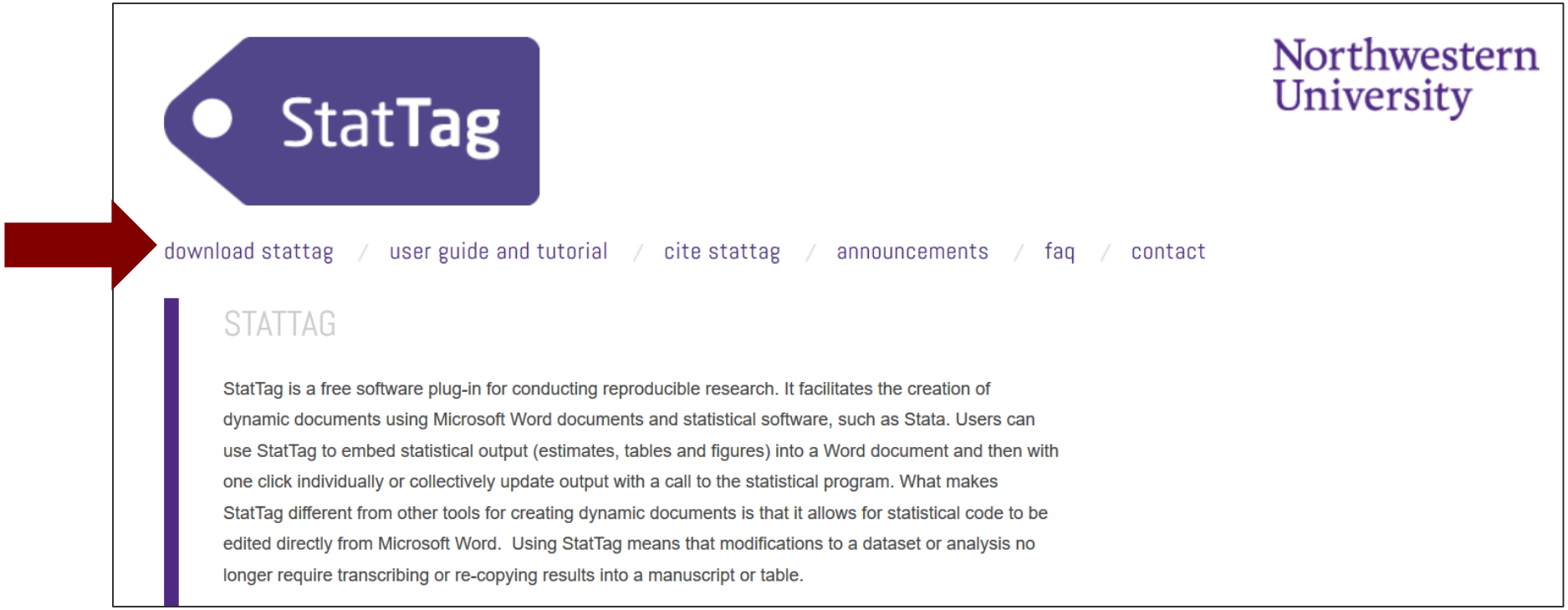
- StatTag *doesn't* store a copy of your data.
- StatTag will eventually store a *read-only* copy of your code

Getting to know StatTag

How do I get StatTag?

StatTag

stattag.org



Getting to know StatTag

How do I cite StatTag?

- We ask that anyone who uses StatTag as a part of their manuscript preparation cite StatTag:
  - Welty, L.J., Rasmussen, L.V., & Baldridge, A.S. (2016). *StatTag*. Chicago, Illinois, United States: Galter Health Sciences Library. doi:10.18131/G3K76

- StatTag was developed with funding through a Clinical Translational Sciences Award (CTSA) to Northwestern University. Tracking the impact of the award is a key metric in demonstrating effectiveness.

- StatTag is distributed under the MIT License

Northwestern Medicine

Summary

Good

Well commented statistical programs, with log files or other record of execution

REDCap or scripted data capture, version control

Systems for linking final manuscript to data, programs, and code

Packages, systems, and workflows that bundle data and programs

Summary

Good	Not so good
Well commented statistical programs, with log files or other record of execution	Analyses conducted on the command line with no record of sequence of code
REDCap or scripted data capture, version control	Data stored in Excel, without record of updates or corrections
Systems for linking final manuscript to data, programs, and code	Published papers with no record of final analyses or data used in manuscript
Packages, systems, and workflows that bundle data and programs	Data and programs unavailable to investigator, reviewers, or colleagues for replication or review

Summary

Good	Not so good
Well commented statistical programs, with log files or other record of execution	Analyses conducted on the command line with no record of sequence of code
REDCap or scripted data capture, version control	Data stored in Excel, without record of updates or corrections
Systems for linking final manuscript to data, programs, and code	Published papers with no record of final analyses or data used in manuscript
Packages, systems, and workflows that bundle data and programs	Data and programs unavailable to investigator, reviewers, or colleagues for replication or review

There are excellent and accessible alternatives to Excel.
There are many advantages to using them.

Acknowledgements

- StatTag Team
  - Leah J. Welty, Biostatistician
  - Luke V. Rasmussen, Lead Software Developer
  - Abigail S. Baldridge, Senior Statistical Analyst
  - Eric W. Whitley, Software Developer

- Development of StatTag was supported, in part, by the National Institutes of Health's National Center for Advancing Translational Sciences, Grant Number UL1TR001422. *The content is solely the responsibility of the developers and does not necessarily represent the official views of the National Institutes of Health.*

Acknowledgements (continued)

- StatTag was inspired in part by the Stata Automation Report project: Lo Magno, G.L. (2013). Sar: Automatic generation of statistical reports using Stata and Microsoft Word for Windows. *The Stata Journal*, 13(1); 39-64.

- StatTag makes use of the following open source projects:

  Scintilla - http://www.scintilla.org/

  ScintillaNET - https://github.com/jacobslusser/ScintillaNET

  Json.NET - http://www.newtonsoft.com/json

Use of these projects does not imply endorsement of StatTag by the respective project owners, or endorsement of the use of these projects by Northwestern University.

References

- Peng, R "Reproducible Research in Computational Science" (Science) 2 December 2011, vol. 334
- Leek and Peng "Opinion: Reproducible Research can still be wrong: Adopting a prevention approach" PNAS, February 10, 2015, vol. 112, no. 6, 1645–1646
- Matthias Schwab, Martin Karrenback, and Jon Claerbout "Making scientific computations reproducible" (2000) *Computing in Science and Engineering*, 2, pp. 61 – 67.
- Roger D. Peng. "Reproducible research and *Biostatistics."* (2009) *Biostatistics,* pp. 405-408.
- Paul Thompson and Andrew Burnett. "Reproducible Research" CORE Issues in Professional and Research Ethics, Volume 1, Paper 6, 2012. Accessed from http://nationalethicscenter.org/content/article/175
- Jonathan Buckheit and David Donoho. "WaveLab and Reproducible Research." (1995) Technical Report No. 474, Department of Statistics, Stanford Univeristy. Accessed from http://statistics.stanford.edu/~ckirby/techreports/NSF/EFS%20NSF%20474.pdf, February 2013.
- Babel: Introduction. OrgMode; Available from: http://orgmode.org/worg/org-contrib/babel/intro.html.
- Buckheit JB, Donoho DL. WaveLab and Reproducible Research. In: Antoniadis A, Oppenheim G, editors. Wavelets and Statistics. Springer New York; 1995. p. 55-81.
- Laine C, Goodman SN, Griswold ME, Sox HC. Reproducible research: moving toward research the public can really trust. Ann Intern Med. 2007;146(6):450-3.
- Thompson PA, Burnett A. Reproducible Research. CORE Issues [serial on the Internet]. 2012; Vol. 1, Paper 6: Available from: https://nationalethicscenter.org/content/article/175.
- Ware J. Reproducible Research Standards and Definitions. CTSPedia; 2010; Available from: http://www.ctspedia.org/do/view/CTSpedia/ReproducibleResearchStandards.
- Baggerly KA, Coombes KR. Deriving chemosensitivity from cell lines: Forensic bioinformatics and reproducible research in high-throughput biology. Annals of Applied Statistics. 2009;3(4):1309-34.
- Baggerly KA, Coombes KR. What information should be required to support clinical "omics" publications? Clin Chem. 2011;57(5):688-90.
- Coombes KR, Wang J, Baggerly KA. Microarrays: Retracing steps [4]. Nat Med. 2007;13(11):1276-7.
- Donoho DL. An invitation to reproducible computational research. Biostatistics. 2010;11(3):385-8.
- Gentleman R. Reproducible research: a bioinformatics case study. Stat Appl Genet Mol Biol. 2005;4:Article2.
- Announcement: Reducing our Irreproducibility. Nature. 2013;496:398 (Editorial).
- Code share: Papers in Nature journals should make computer code accessible where possible. Nature. 2014;514:536 (Editorial).
- Diggle PJ, Zeger SL. Embracing the concept of reproducible research. Biostatistics (Oxford, England). 2010;11(3):375.
- Godlee F, Groves T. The new BMJ policy on sharing data from drug and device trials. BMJ (Clinical research ed). 2012;345:e7888.
- Groves T, Godlee F. Open science and reproducible research. BMJ (Clinical research ed). 2012;344:e4383.
- McNutt M. Journals unite for reproducibility. Science (New York, NY). 2014;346(6210):679.
- Peng RD. Reproducible research and Biostatistics. Biostatistics (Oxford, England). 2009;10(3):405-8.

Questions?

Thank You!