

Data Formatting Issues

Managing Your Research Data— Best practices in Research Data Management for the Biological Sciences



CANCER
RESEARCH
UK

CAMBRIDGE
INSTITUTE

Reproducible Research

- At some point in the future, someone, somewhere, might want to repeat your analysis for themselves or re-use your data.
 - which will most likely be ***you!***
- Assuming that you'll be able to remember all the steps involved is dangerous, so making sure that everything is well-documented is key.
- The documentation involves not only the methods used, but the files used as input and any transformations performed on them.



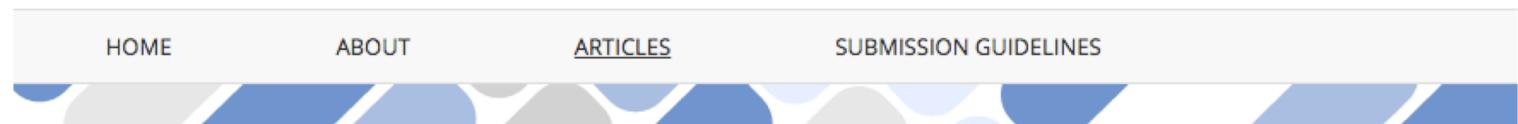
CANCER
RESEARCH
UK

CAMBRIDGE
INSTITUTE

Five selfish reasons

- Florian Markowetz has a great talk on why we should work reproducibly
- There is a [Genome Biology paper](#) that you should read.

Genome Biology



COMMENT | OPEN ACCESS

Five selfish reasons to work reproducibly

Florian Markowetz 

Genome Biology 2015 16:274 | DOI: 10.1186/s13059-015-0850-7 | © Markowetz. 2015

Published: 8 December 2015

Abstract

And so, my fellow scientists: ask not what you can do for reproducibility; ask what reproducibility can do for you! Here, I present five reasons why working reproducibly pays off in the long run and is in the self-interest of

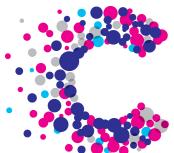


CANCER
RESEARCH
UK

CAMBRIDGE
INSTITUTE

A famous example

- Probably the most (in)famous example of failure to reproduce a study, which actually *put people's lives at risk* and rallied statisticians into action
- Keith Baggerly's lecture on the scandal is a ***must-see***.
<https://www.youtube.com/watch?v=7gYls7uYbMo>
- If that wasn't enough to give you sleepless nights – Visit <http://retractionwatch.com>



CANCER
RESEARCH
UK

CAMBRIDGE
INSTITUTE

Are spreadsheets programs like Excel evil?

-Not necessarily.

- Often much more convenient to eye-ball a spreadsheet and get an overall impression of your data.

- But they have *limitations* making them not ideal for large-scale analyses.

- Doing things by-hand only invites you to make copy-and-paste errors etc.

- R cannot read all files as if by magic



CANCER
RESEARCH
UK

CAMBRIDGE
INSTITUTE

Helpful Data Validation features in Excel

- Excel data validation feature
- Select a column
 - In the menu bar, choose “Data”
 - Validation
- Integer or decimal number
- Range
- List of possible values
- Limited length text



CANCER
RESEARCH
UK

CAMBRIDGE
INSTITUTE

Less helpful features in Excel

- When identifiers are long integers
 - $1000000 = 1e06$
 - [Issue with Illumina microarray chip IDs](#)
- [Excel can convert gene names to dates](#)
 - SEPT2 (Septin 2) → '2-Sep'
 - MARCH1 (Membrane-Associated Ring Finger (C3HC4) 1, E3 Ubiquitin Protein Ligase) → '1-Mar'
- Conversion of Study ID codes! ☹



CANCER
RESEARCH
UK

CAMBRIDGE
INSTITUTE

Data Handling rules:

In TV's NCIS, Special Agent Leroy 'Jethro' Gibbs has a set of rules. Failure to observe Gibbs Rules results in a 'Gibbs Slap'

We're not that cruel but here are some rules to follow when dealing with your data.

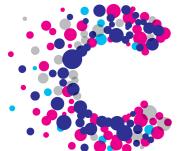


Image credit:
<https://saisoto.deviantart.com/art/Gibbs-Slap-173459786>

Rule 1 -Never work directly on the raw data



<http://www.inquisitr.com/309687/jesus-painting-restoration-goes-wrong-well-intentioned-old-lady-destroys-100-year-old-fresco/>



CANCER
RESEARCH
UK

CAMBRIDGE
INSTITUTE

Rule 1 -Never work directly on the raw data

- Hard to reverse all the manual steps performed and invites errors
- Store the original data somewhere **safe**
 - see later on today (OpenRefine exercise)



CANCER
RESEARCH
UK

CAMBRIDGE
INSTITUTE

Example 1 – How many inconsistencies can you spot?

Patient ID	Sex	Date of Diagnosis	Tumour Size
1	M	01-01-2013	3.1
2	f	04-18-1998	1.5
3	Male	1st of April 2004	105
4	Female	NA	67
5	F	2010/03/12	4.2
6	F		3.6
7	M	1994-11-05T08:15:30-05:00	232

Rule 2 - Maintain consistency

How many ways can you say “female”?

18-day pregnant females	female (lactating)	individual female	worker caste (female)
2 yr old female	female (pregnant)	lgb*cc females	sex: female
400 yr. old female	female (outbred)	mare	female, other
adult female	female parent	female (worker)	female child
asexual female	female plant	monosex female	femal
castrate female	female with eggs	ovigerous female	3 female
cf.female	female worker	oviparous sexual females	female (phenotype)
cystocarpic female	female, 6-8 weeks old	worker bee	female mice
dikaryon	female, virgin	female enriched	female, spayed
dioecious female	female, worker	pseudohermaphroditic female	femlale
diploid female	female(gynoecious)	remale	metafemale
f	femele	semi-engorged female	sterile female
famale	female, pooled	sexual oviparous female	normal female
femail	femalen	sterile female worker	sf
female	females	strictly female	vitellogenic replete female
female - worker	females only	tetraploid female	worker
female (alate sexual)	gynoecious	thelytoky	hexaploid female
female (calf)	healthy female	female (gynoecious)	female (f-o)
hen	probably female (based on morphology)		

female (note: this sample was originally provided as a \"male\" sample to us and therefore labeled this way in the brawand et al. paper and original geo submission; however, detailed data analyses carried out in the meantime clearly show that this sample stems from a female individual)",



CANCER
RESEARCH
UK

CAMBRIDGE
INSTITUTE

Courtesy of N. Silvester, European Nucleotide Archive, EMBL-EBI

Example 1 – applying Rule 2

- Consistency: F, female, f, fem, 2, ...
- Units
 - cm or mm; days, months or years
- You can introduce inconsistencies without realising it
 - blank spaces (whitespace) at the end of text
 - "Male " is not the same as "Male"
- Document choices you make about units in a *README* file



CANCER
RESEARCH
UK

CAMBRIDGE
INSTITUTE

Regarding dates

credit: @myusuf3



CANCER
RESEARCH
UK

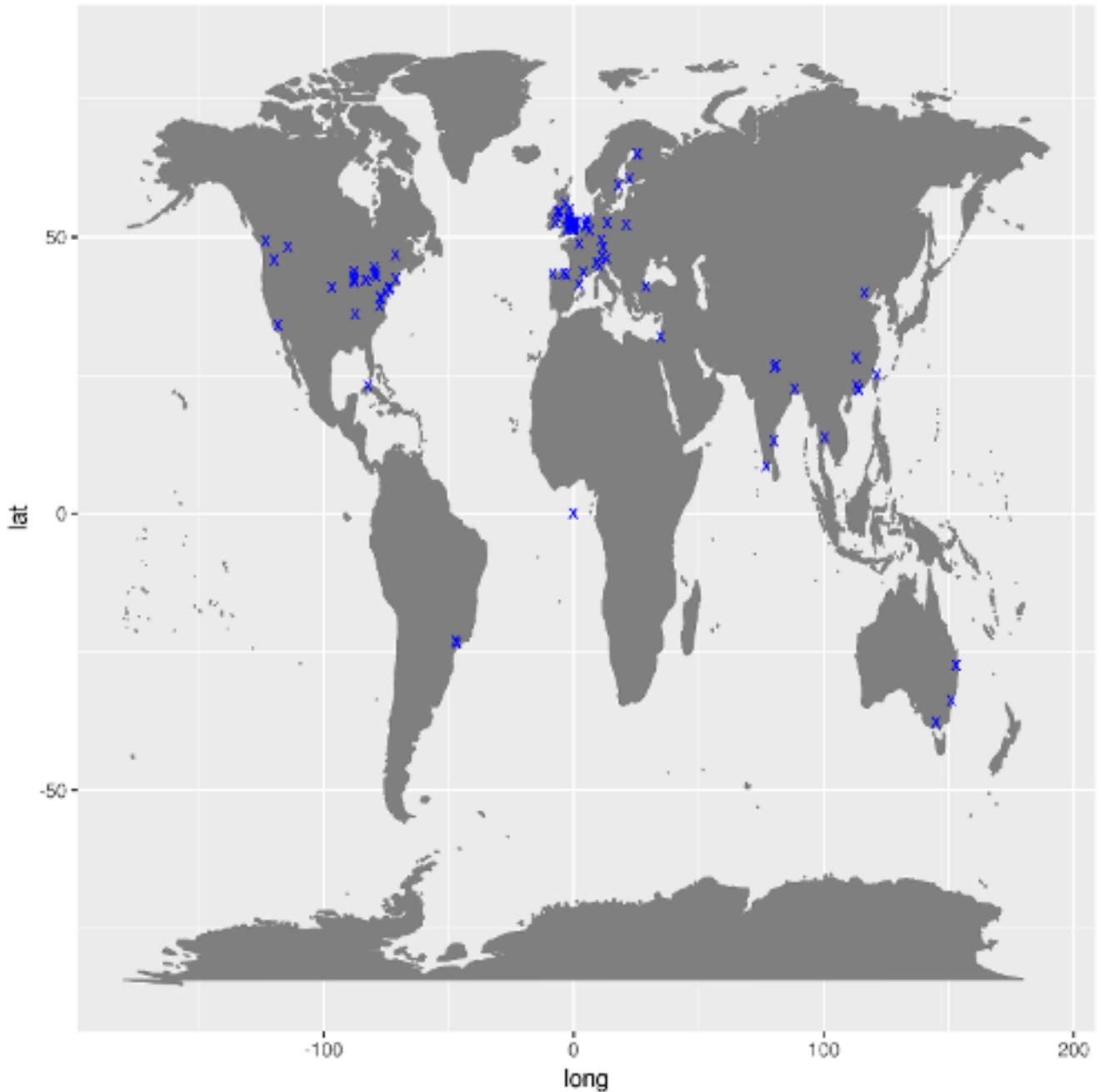
CAMBRIDGE
INSTITUTE

Example 1 – corrected using Rule 2

Patient ID	Sex	Date of Diagnosis	Tumour Size
001	M	2013-01-01	3.1
002	F	1998-04-18	1.5
003	M	2004-04-01	1.05
004	F	NA	0.67
005	F	2010-03-12	4.2
006	F	NA	3.6
007	M	1994-11-05	2.32

Rule 3 – Missing values

Figure showing locations of visitors to my Prostate Cancer
[data portal](#)



CANCER
RESEARCH
UK

CAMBRIDGE
INSTITUTE

Rule 3 - Don't use 0 to mean missing

- Zero values are data!
 - Sometimes extreme values such as 999 are sometimes used
- NA is Ok, but what if NA is a valid category in your data?
 - R will recognise NA as a missing value and can ignore it in calculations
- Safest to leave the cell *empty*
 - but you need to be careful with blank spaces



CANCER
RESEARCH
UK

CAMBRIDGE
INSTITUTE

Rule 4 - Fill in all the cells

Example 2

Patient ID	Date	Value
1	2015-06-14	213
2		76.5
3	2015-06-18	32
4		120.3
5		109
6	2015-06-20	
7		143



CANCER
RESEARCH
UK

CAMBRIDGE
INSTITUTE

Rule 4 - Fill in all the cells

- It is tempting to make the table look cleaner by not repeating some values
- Fill in all cells!
 - otherwise, problems when sorting
- Empty cell:
 - missing value?
 - value meant to be repeated multiple times?
- Make sure it's clear that the data is missing and not unintentionally left blank



CANCER
RESEARCH
UK

CAMBRIDGE
INSTITUTE

Example 2 Corrected using Rule 4

Patient ID	Date	Value
1	2015-06-14	213
2	2015-06-14	76.5
3	2015-06-18	32
4	2015-06-18	120.3
5	2015-06-18	109
6	2015-06-20	
7	2015-06-20	143



CANCER
RESEARCH
UK

CAMBRIDGE
INSTITUTE

Rule 5 - Make it rectangular

- The computer expects a very rigid shape of data with rows and columns
- Each column is a *variable* being examined
- Each row is an observation
- A concept commonly known as *tidy data*



CANCER
RESEARCH
UK

CAMBRIDGE
INSTITUTE

Rule 5 - Make it rectangular

	A	B	C	D	E	F	G	H	I
1		1 min				5 min			
2	strain	normal		mutant		normal		mutant	
3	A	147	139	166	179	334	354	451	474
4	B	246	240	178	172	514	611	412	447



CANCER
RESEARCH
UK

CAMBRIDGE
INSTITUTE

Rule 5 - Make it rectangular

	A	B	C	D	E
1	strain	genotype	min	replicate	response
2	A	normal	1	1	147
3	A	normal	1	2	139
4	B	normal	1	1	246
5	B	normal	1	2	240
6	A	mutant	1	1	166
7	A	mutant	1	2	179
8	B	mutant	1	1	178
9	B	mutant	1	2	172
10	A	normal	5	1	334
11	A	normal	5	2	354
12	B	normal	5	1	514
13	B	normal	5	2	611
14	A	mutant	5	1	451
15	A	mutant	5	2	474
16	B	mutant	5	1	412
17	B	mutant	5	2	447



CANCER
RESEARCH
UK

CAMBRIDGE
INSTITUTE

More

- Don't put too much information in one cell
 - 1 cell = 1 piece of information
- Don't include units such as "30 g" → "g" in the column name
 - <http://unitsofmeasure.org/ucum.html>
- Write notes in a separate column or data dictionary or metadata
 - "0 (below threshold)"



CANCER
RESEARCH
UK

CAMBRIDGE
INSTITUTE

More

- Don't put too much information in one cell
 - 1 cell = 1 piece of information
- Don't include units such as "30 g" → "g" in the column name
 - <http://unitsofmeasure.org/ucum.html>
- Write notes in a separate column or data dictionary or metadata
- "0 (below threshold)"

Design for machine readability

- NO calculations
- NO font **colours**
- NO highlighting

Computer doesn't need or recognize it!



CANCER
RESEARCH
UK

CAMBRIDGE
INSTITUTE

Recapping the ‘Rules’

Rule 1 -Never work directly on the raw data

Rule 2 - Maintain consistency

Rule 3 - Don't use 0 to mean missing

Rule 4 - Fill in all the cells

Rule 5 - Make it rectangular



CANCER
RESEARCH
UK

CAMBRIDGE
INSTITUTE

Write Protection

Mac

- Right click on the file in Finder
- Select “Get Info”
- Sharing and permission
- Privilege
- Read only

Write Protection

Windows

- Right click on the file in Windows Explorer
- Properties
- General tab
- Attributes
- Select the box for “read only”

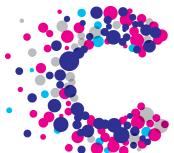


CANCER
RESEARCH
UK

CAMBRIDGE
INSTITUTE

Practical - Practice (makes perfect)

- Look at the file [patient-data.txt](#)
 - a simulated, but representative, example of ***bad data***
 - discuss with your neighbours (around 5 minutes)
- The next step is to look at how to clean the data with *Open Refine*
NB If you are planning to use your own laptop then you will need to install the free package ‘OpenRefine’ from <http://openrefine.org> .



CANCER
RESEARCH
UK

CAMBRIDGE
INSTITUTE

Practical - Project proposals looking for a data plan (DataLine?)

(Discuss these in pairs – what considerations do you think a Funder needs?)

- Consented patient tumour samples will be analysed for particular SNPs. Sequencing will be done externally with computational analysis carried out in-house. A paper will be written and the experimental data shared.
- GP practices will get patients to fill in questionnaires relating to lifestyle and diet and to provide biological samples to be analysed at researchers lab for bio-markers of cardiac disease.
- An archaeological dig is to be organised in North Walsham where it is anticipated that Roman remains will be uncovered. Finds will be verified by other national experts, site data will be stored for use by other academics and the findings made publicly available.
- A trial is planned of a new anti-depressant to help treat patients. One group will be given the new drug in their combination therapy and the other group will receive an existing one. A comparative analysis of transcribed patient interviews will be undertaken.



CANCER
RESEARCH
UK

CAMBRIDGE
INSTITUTE