Reflections from a R Lady in the Untidyverse

Hilary A. Robbins

Scientist, International Agency for Research on Cancer

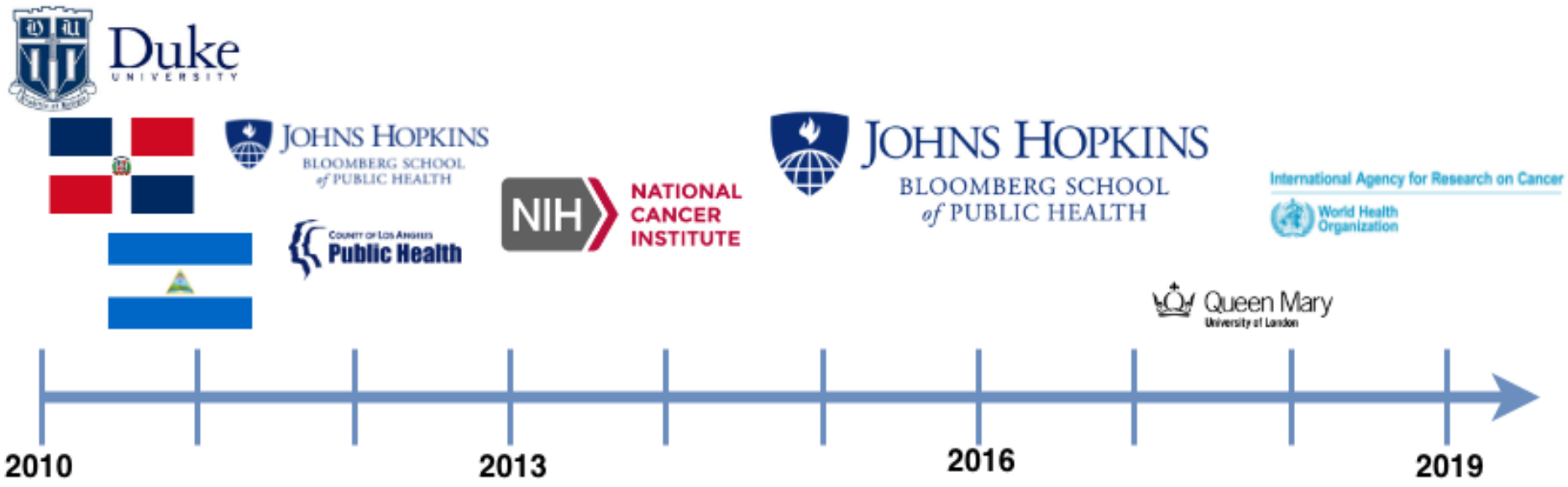
15 April 2019

RobbinsH@iarc.fr
@hilaryarobbins

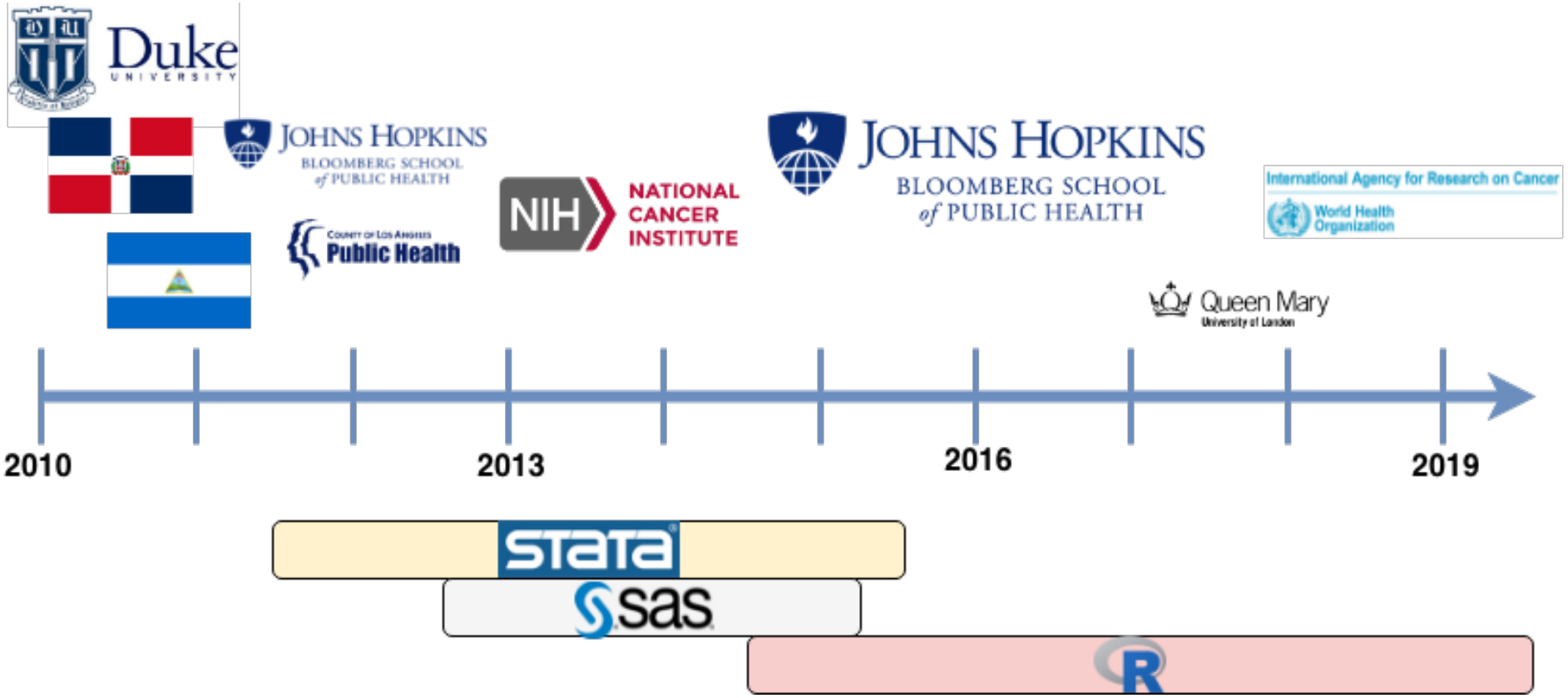
Overview

About Me

Who Am I?

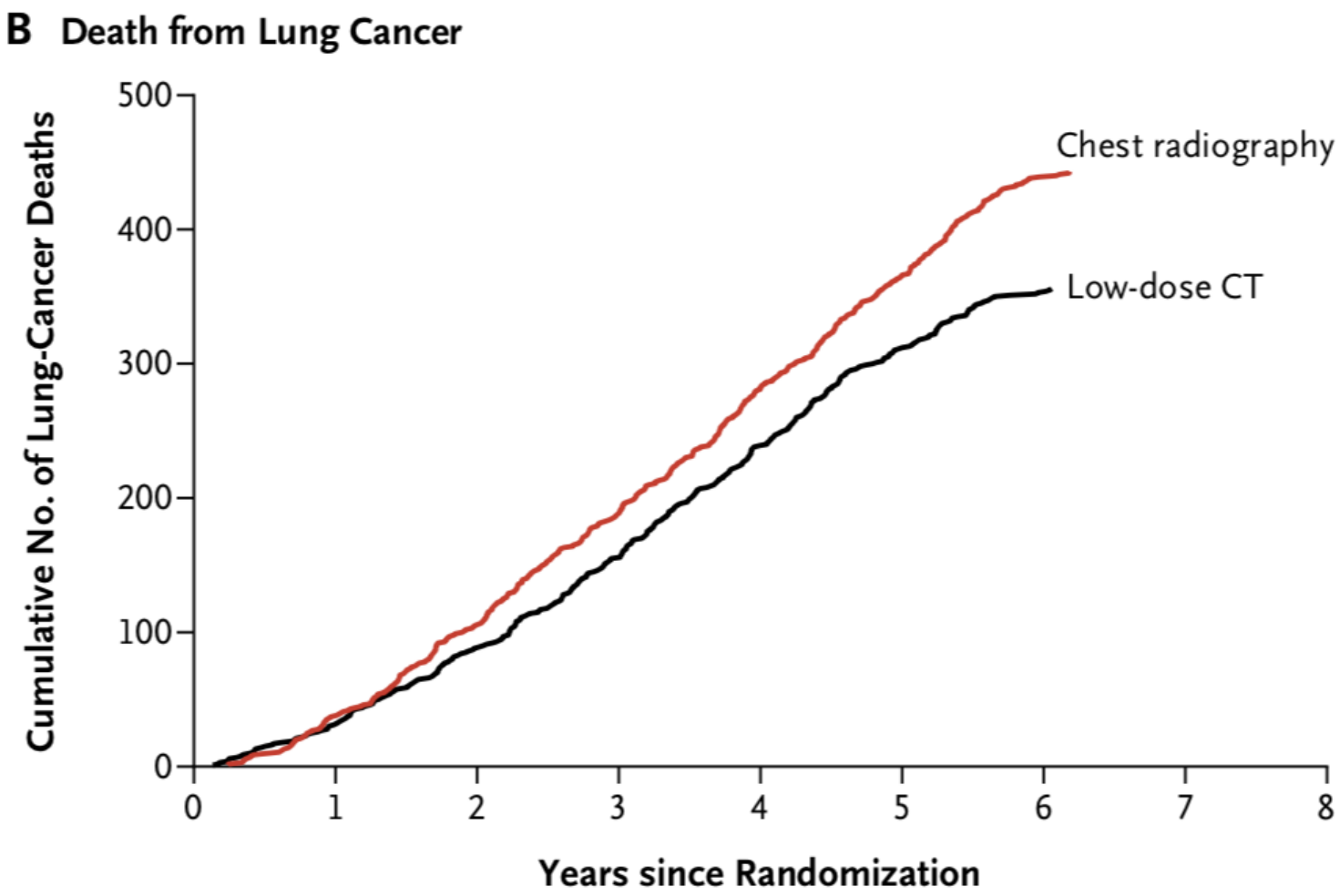

Duke
UNIVERSITY
JOHNS HOPKINS
BLOOMBERG SCHOOL
of PUBLIC HEALTH
County of Los Angeles
Public Health
NIH
NATIONAL
CANCER
INSTITUTE
JOHNS HOPKINS
BLOOMBERG SCHOOL
of PUBLIC HEALTH
International Agency for Research on Cancer
World Health
Organization
Queen Mary
University of London
2010
2013
2016
2019

Who am I?


Duke
UNIVERSITY
JOHNS HOPKINS
BLOOMBERG SCHOOL
of PUBLIC HEALTH
County of Los Angeles
Public Health
NIH
NATIONAL
CANCER
INSTITUTE
JOHNS HOPKINS
BLOOMBERG SCHOOL
of PUBLIC HEALTH
International Agency for Research on Cancer
World Health
Organization
Queen Mary
University of London
2010
2013
2016
2019
STATA
sas
R

Benefits and Harms of Lung Cancer Screening (or, How I Ended up Making an Infographic)

CT screening reduces lung cancer mortality



National Lung Screening Trial (NEJM 2011)

The National Health Service (NHS) plans to roll out lung cancer screening across England

Home News Publications Statistics Blogs Events Contact us

NHS England

Search

About NHS England Our work Commissioning Get involved

Search news

You can use the filters to show only news items that match your interests

Keyword

Topic

Select topic

News

NHS to rollout lung cancer scanning trucks across the country

8 February 2019

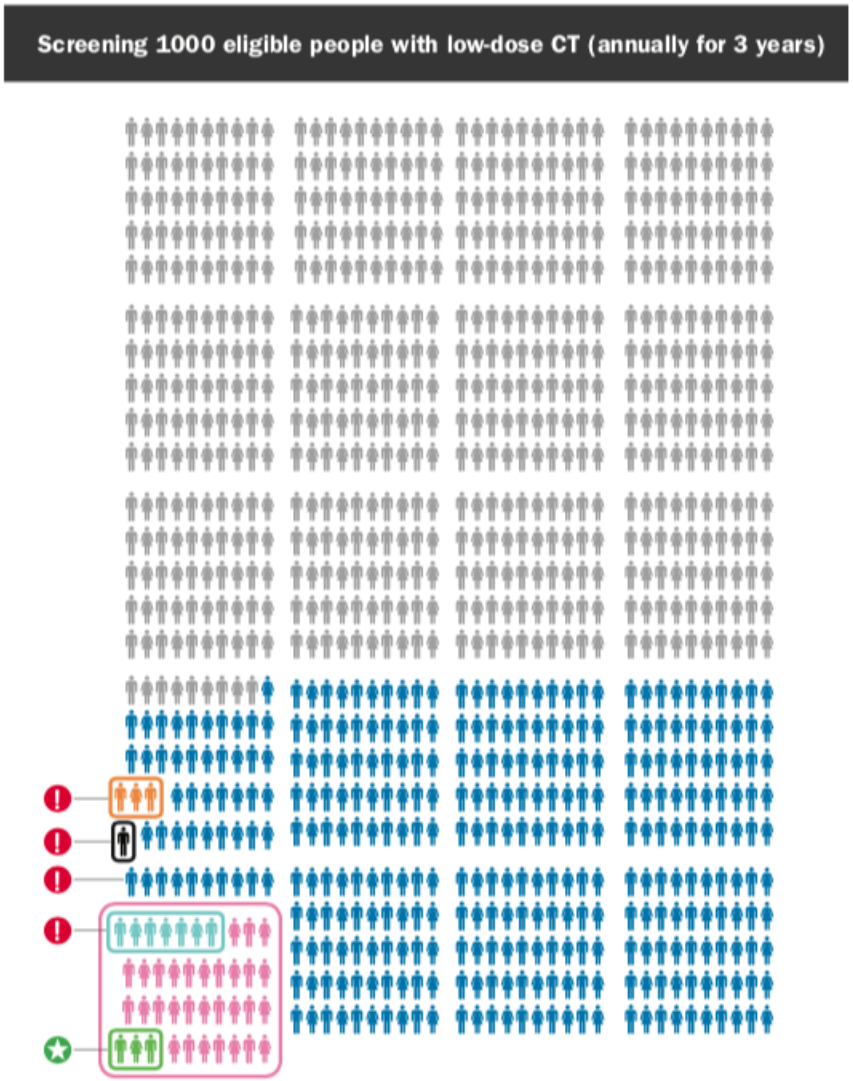
Cancer Respiratory

Lung cancer scanning trucks that operate from supermarket car parks are being rolled out across the country in a drive to save lives by catching the condition early, NHS England announced today.

Due to incorrect interpretations of data, misinformation has spread in the public

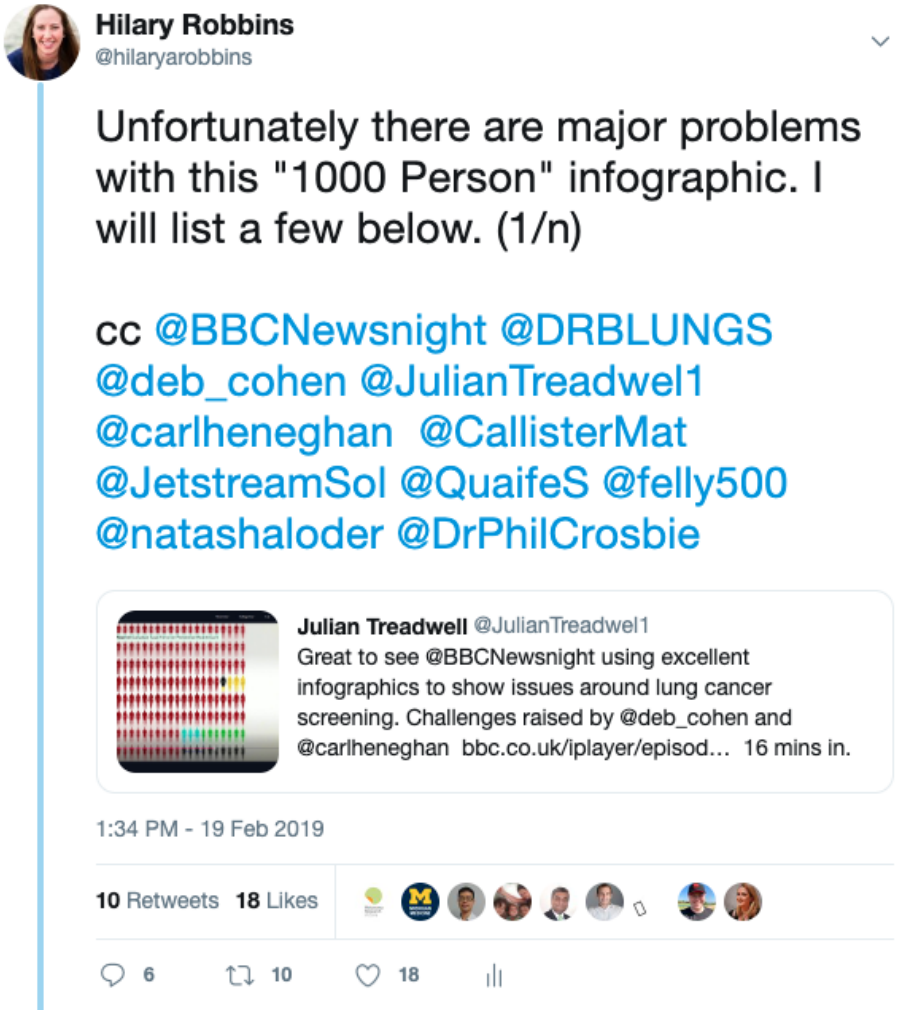
BBC Newsnight, 11 February 2019

Due to incorrect interpretations of data, misinformation has spread in the public



609	will have a negative low-dose CT scan result	
40	will be diagnosed with lung cancer	
351	will have a positive scan result and find out after further testing that they do not have cancer (false positive)	Harm
7	of the 40 diagnosed lung cancers would not have caused illness or death (overdiagnosis)	Harm
3	will have major complications from invasive follow-up tests	Harm
1	will die from invasive follow-up testing	Harm
3	fewer people will die from lung cancer (vs. when screening with chest x-ray)	Benefit

It all started... on Twitter



Data to the rescue: What if an old study had used a new protocol?

- There would have been:
  - Fewer false-positive results
  - Fewer invasive diagnostic procedures
  - Slightly fewer lives saved (some cancers would be missed)
- We can calculate these outcomes by reclassifying screen results based on the new protocol.

The Data

	pid	case	age	female	edu6	bmi	cpd	smkyears	truefalse_scrnres_ly0	truefalse_scrnres_ly1	truefalse_scrnres_ly2
1	100002	3	66	0	2	26.60575	20	52	4	4	4
2	100004	0	60	0	4	29.41122	40	17	3	3	4
3	100005	0	64	0	1	34.45311	40	46	3	3	3
4	100009	0	55	0	6	37.11892	30	35	4	4	4
5	100010	0	68	0	4	30.40657	40	42	4	4	4
6	100011	0	57	1	4	25.12497	60	22	NA	NA	NA
7	100012	1	61	1	6	22.23791	20	37	2	1	NA
8	100014	0	55	1	4	23.95760	30	40	4	4	4
9	100015	0	59	0	4	29.79844	30	50	4	4	4
10	100019	0	61	0	4	23.96024	40	39	4	3	4
11	100020	0	58	0	3	29.29167	40	39	4	4	4
12	100023	0	62	0	2	23.05363	20	43	4	4	NA
13	100024	0	60	0	3	33.96135	20	42	4	4	4
14	100026	0	57	0	3	35.14303	30	41	3	3	4
15	100029	0	56	1	1	24.40972	20	46	4	4	NA
16	100030	0	60	1	5	27.48392	20	35	4	4	4
17	100031	0	64	1	5	26.49673	16	46	4	4	4
18	100032	0	58	0	6	24.40488	20	40	4	4	4
19	100035	0	55	1	3	22.09429	20	38	3	3	4
20	100037	0	56	0	4	26.30965	40	41	4	4	4
21	100040	0	60	0	6	26.30601	30	42	4	3	4
22	100041	0	68	0	4	26.30965	20	41	4	4	4

Dplyr Basics: Verbs

```
filter()      # select rows
select()      # select columns
arrange()     # sort/reorder rows
mutate()      # add new variables
group_by()    # divide rows into groups
summarise()   # calculate summary statistics
```

Introduction to Dplyr

Dplyr Basics: Why use Piping?

```
filter(
    summarise(
        select(
                group_by(flights, year, month, day),
        arr_delay, dep_delay),
    arr = mean(arr_delay, na.rm = TRUE),
    dep = mean(dep_delay, na.rm = TRUE)
    ),
arr > 30 | dep > 30
)
```

CRAN - Introduction to Dplyr

Dplyr Basics: Why use Piping?

```
flights %>%
  group_by(year, month, day) %>%
  select(arr_delay, dep_delay) %>%
  summarise(arr = mean(arr_delay, na.rm = TRUE),
            dep = mean(dep_delay, na.rm = TRUE)) %>%
  filter(arr > 30 | dep > 30)
```

CRAN - Introduction to Dplyr

Dplyr Basics: Why use Piping?

```
   year month   day   arr   dep
  <int> <int> <int> <dbl> <dbl>
1  2013     1    16  34.2  24.6
2  2013     1    31  32.6  28.7
3  2013     2    11  36.3  39.1
4  2013     2    27  31.3  37.8
5  2013     3     8  85.9  83.5
6  2013     3    18  41.3  30.1
7  2013     4    10  38.4  33.0
8  2013     4    12  36.0  34.8
9  2013     4    18  36.0  34.9
10 2013     4    19  47.9  46.1
# ... with 39 more rows
```

The Data

	pid	STUDY_YR	SCT_AB_DESC	SCT_PRE_ATT	SCT_LONG_DIA	SCT_EPI_LOC	SCT_MARGINS	sct_ab_preExist
1	100002	0	65	NA	NA	NA	NA	NA
2	100002	1	64	NA	NA	NA	NA	NA
3	100002	2	65	NA	NA	NA	NA	NA
4	100004	0	51	1	4	1	2	NA
5	100004	0	64	NA	NA	NA	NA	NA
6	100004	0	65	NA	NA	NA	NA	NA
7	100004	1	51	1	4	1	2	2
8	100004	1	65	NA	NA	NA	NA	NA
9	100004	2	52	NA	NA	NA	NA	NA
10	100005	0	51	1	6	1	2	NA
11	100005	0	52	NA	NA	NA	NA	NA
12	100005	0	59	NA	NA	NA	NA	NA
13	100005	0	60	NA	NA	NA	NA	NA
14	100005	1	51	1	6	1	2	2
15	100005	1	60	NA	NA	NA	NA	NA
16	100005	1	52	NA	NA	NA	NA	NA
17	100005	1	53	NA	NA	NA	NA	NA
18	100005	1	59	NA	NA	NA	NA	NA
19	100005	2	51	1	6	1	2	2
20	100005	2	65	NA	NA	NA	NA	NA
21	100005	2	56	NA	NA	NA	NA	NA

Using Dplyr to Look at Data

```
abnormalities %>%
    filter(STUDY_YR==0 & SCT_AB_DESC==51) %>%
    select(SCT_PRE_ATT, SCT_LONG_DIA,
        SCT_EPI_LOC, SCT_MARGINS) %>%
    sample_n(30) %>%
    View()
```

Using Dplyr to Look at Data

	SCT_PRE_ATT	SCT_LONG_DIA	SCT_EPI_LOC	SCT_MARGINS
6346	2	8	3	3
2648	2	5	4	3
8287	2	17	1	1
7831	1	6	6	2
2211	1	5	6	2
8977	1	5	2	2
6050	2	10	1	3
9338	1	6	3	2
7744	1	5	5	2
7261	1	10	1	2
7034	2	18	1	3
2776	2	11	3	3
587	1	9	6	1
6847	1	5	6	2
2632	1	6	2	2
306	2	4	6	3
8523	1	4	6	2
7867	1	7	3	2
6384	1	6	4	2
7447	1	6	4	3
3087	1	6	3	2
9164	1	6	2	2

Using Dplyr to Manipulate Data

```
abnormalities.person.level <-
    abnormalities %>%
    group_by(pid, STUDY_YR) %>%
    summarise(longest.diam = max(SCT_LONG_DIA, na.rm=T),
      any.nodule = as.numeric(any(SCT_AB_DESC==51, na.rm=T)),
      emphysema = as.numeric(any(SCT_AB_DESC==59, na.rm=T)))
```

Using Dplyr to Manipulate Data

	pid	STUDY_YR	longest.diam	any.nodule	emphysema
1	100002	0	0	0	0
2	100002	1	0	0	0
3	100002	2	0	0	0
4	100004	0	4	1	0
5	100004	1	4	1	0
6	100004	2	0	0	0
7	100005	0	6	1	1
8	100005	1	6	1	1
9	100005	2	6	1	0
10	100009	1	0	0	0
11	100009	2	0	0	0
12	100010	0	0	0	0
13	100010	2	0	0	0
14	100012	0	8	1	0
15	100012	1	15	1	0
16	100014	0	0	0	0
17	100014	2	0	0	1
18	100015	0	0	0	0
19	100015	1	0	0	0
20	100015	2	0	0	0
21	100019	0	0	0	1
22	100019	1	14	1	1

What did I Learn?

- Data skills create opportunities
- Sometimes, the first step is to realize that you are the person who is best positioned to do something
- It helps to recognize the feeling of uncertainty, so you can lean in to it
- If you consistently build your network, you can draw on it when needed

Being an R lady is not easy

Fewer early opportunities
Unconscious bias
Imposter's syndrome
Biology
We don't live in the tidyverse.

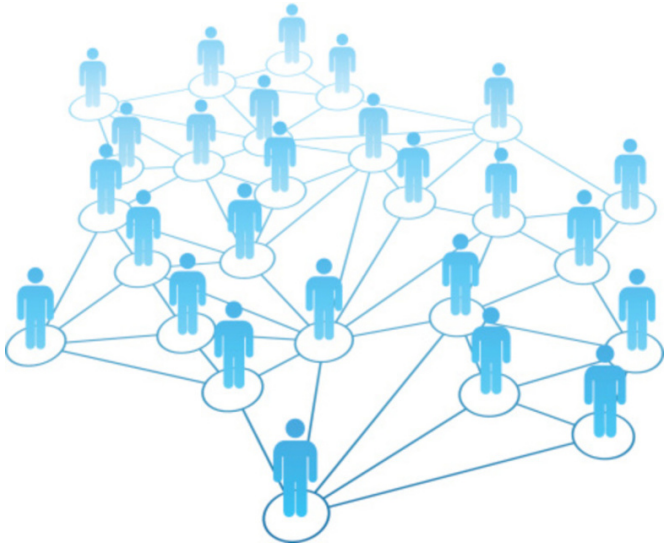
My Toolkit

Know your stuff

but own your limits
and your mistakes

Build your network

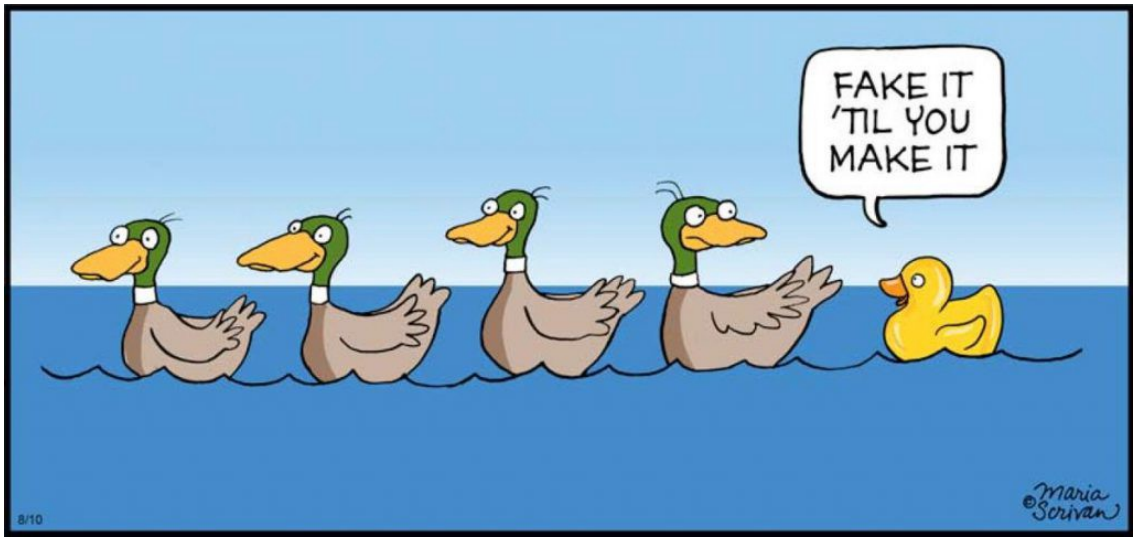
- Quality over quantity
- Work with new people
- Be proactive



Manage how you are perceived by remembering that you matter

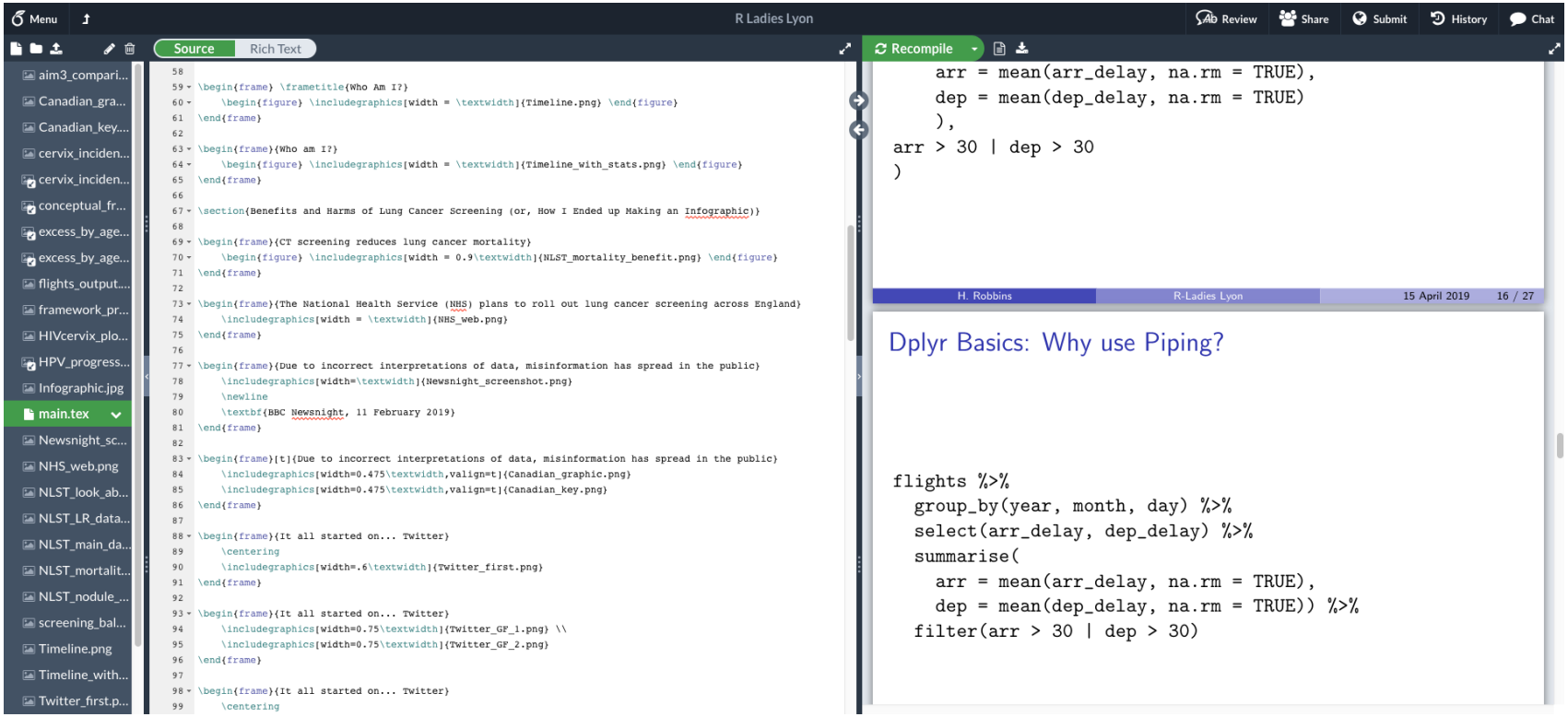
"Sorry, but..."	"Did you control for smoking?"
"I'm doing a project on X"	"My work develops X"
"It was just my master's project"	"It was published in Z"
"I know you're busy, but..."	"Would you like to meet to discuss?"

Fake it till you make it



How do I make slides in LaTeX?

I use the Beamer environment on Overleaf:



[Get started here]

Reflections from a R Lady in the Untidyverse

Hilary A. Robbins

Scientist, International Agency for Research on Cancer

15 April 2019

RobbinsH@iarc.fr
@hilaryarobbins