October, 15th 2017

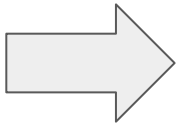
A Day in the Life of Two (Legit) Data Scientists

Sandy Griffith
Senior Methodologist, Flatiron
Elizabeth Sweeney
Quantitative Scientist, Flatiron

flatiron

My background

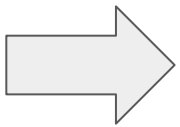
Academic biostatistics



Healthcare tech

My background

Academic biostatistics

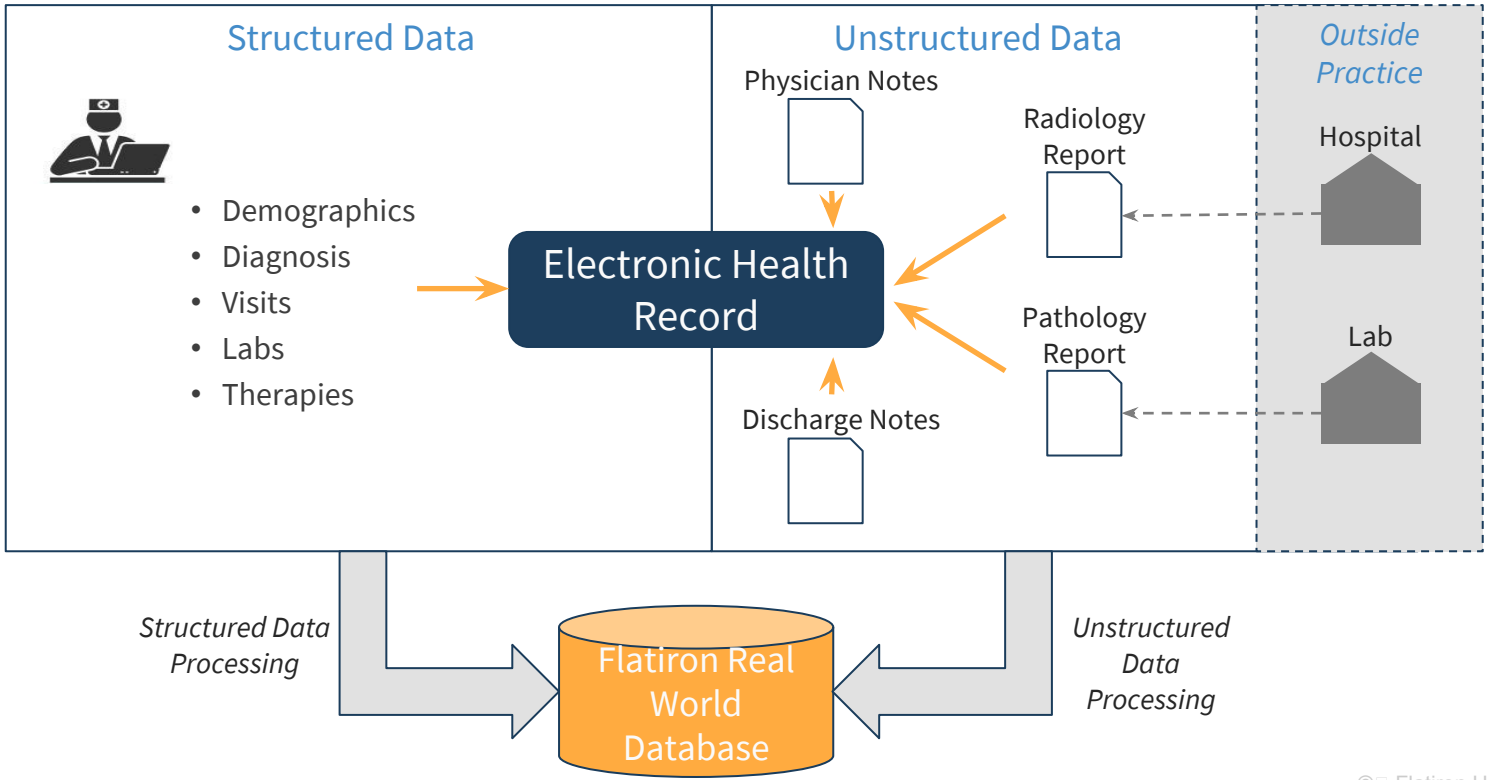


Healthcare tech

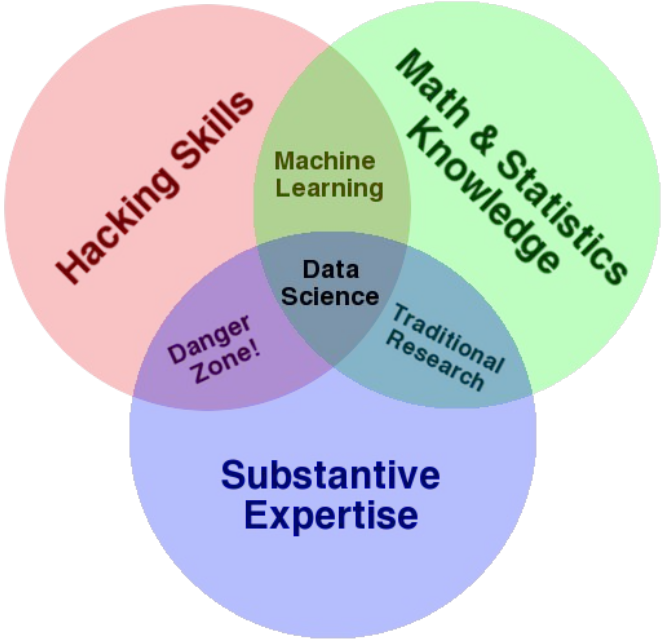
Flatiron's mission is to serve cancer patients and our partners by dramatically improving treatment and accelerating research.

Our Mission

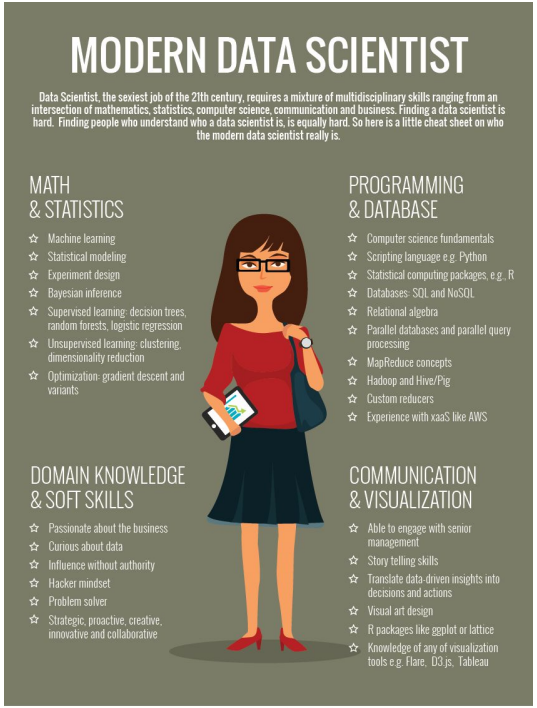
Flatiron processes EHR data at scale



What is a data scientist?



Venn credit: Drew Conway



What is a data scientist *at Flatiron Health*?

At least 3 functional teams that may perform “data science” work:

1. Quantitative Sciences (epidemiologists, biostatisticians, “data scientists”)
2. Data insights engineering (engineers tied closely to products and customers, “data scientists”)
3. Software engineering (engineers, often with computer science background, work on a wide range of software development, “data scientists”)

What types of teams do we work on?

Staffed to cross-functional teams with oncologists, product managers, and many others

Examples of cross-functional product teams:

1. Endpoints development and validation
2. Machine learning platforms
3. Clinical trials

What is the collaborative process like?

Agile software development: set of values and principles for software development that include **adaptive planning**, frequent delivery, **simplicity**, close **daily collaboration**, **rapid and flexible response** to changing requirements[1]

Team “ceremonies”

- Daily standups
- Sprint planning
- Retros
- Blameless postmortems

1. Beck, et al. (2001) "Principles behind the Agile Manifesto"

Sprint planning

- Sprints can vary in length, often 2 weeks
- Create tickets for work, or pull tickets in from the backlog
- Estimate complexity, not time
- Decide what is in scope for the sprint, and what is “below the line”
- Pick a name and start sprinting!

Frequent delivery: What exactly do we deliver?

- Analysis plans and/or design documents
- Code and code reviews
- Lab notebooks
- QA reports
- Analysis results
- Dashboards
- Datasets
- Presentations
- Publications

Focus on the “why”, not the “who”

Blameless PostMortems and a Just Culture

Posted by **John Allspaw** on May 22, 2012

Last week, Owen Thomas wrote a flattering article over at Business Insider on how we handle errors and mistakes at Etsy. I thought I might give some detail on how that actually happens, and why.

Collaboratively compiled onboarding advice

Compiled by and for new team members, especially those working in tech for the first time:

1. Be proactive, vocalize when blocked
2. Submit incomplete work: “30% feedback”
3. Let others know what you do and don’t know
4. Projects are not owned by any single person
5. Plan for interdependencies
6. Simpler is better

Journey to Flatiron Health: Elizabeth Sweeney, PhD

Journey to Flatiron Health: Elizabeth Sweeney, PhD

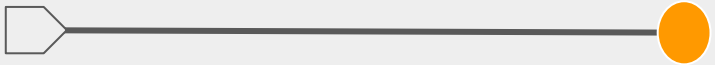
Background: Oncology Endpoints

- Overall Survival (OS) is commonly accepted as primary measure of efficacy of given intervention.
- One (of many) issues with OS is that the follow-up to observe deaths long and expensive.
- Alternative: Progression-Free Survival (PFS) as a *surrogate endpoint* for OS in oncology research.
- PFS is the the length of time during and after the treatment of a disease that a patient lives with the disease but it does not get worse.

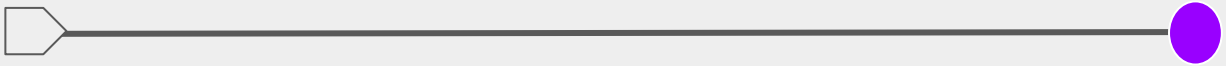
PFS and OS

Underlying true events

PFS



OS



Real World Endpoints

- Work to develop and validate a real-world progression endpoint in advanced non-small cell lung cancer (aNSCLC) derived from electronic medical records (EMRs)

Clinical Trial

Stringent criteria for a progression event (RECIST criteria)

Patients come to the clinical at predetermined time intervals to be assessed for progression with radiologic scans

Real World

Progression is anchored on information in a clinician note

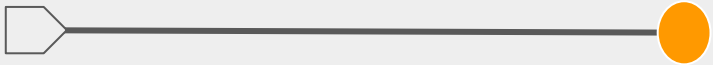
Patients come to the clinical at different times, which can depend on how sick the patient is

Scan timing: Constant Monitoring Setting

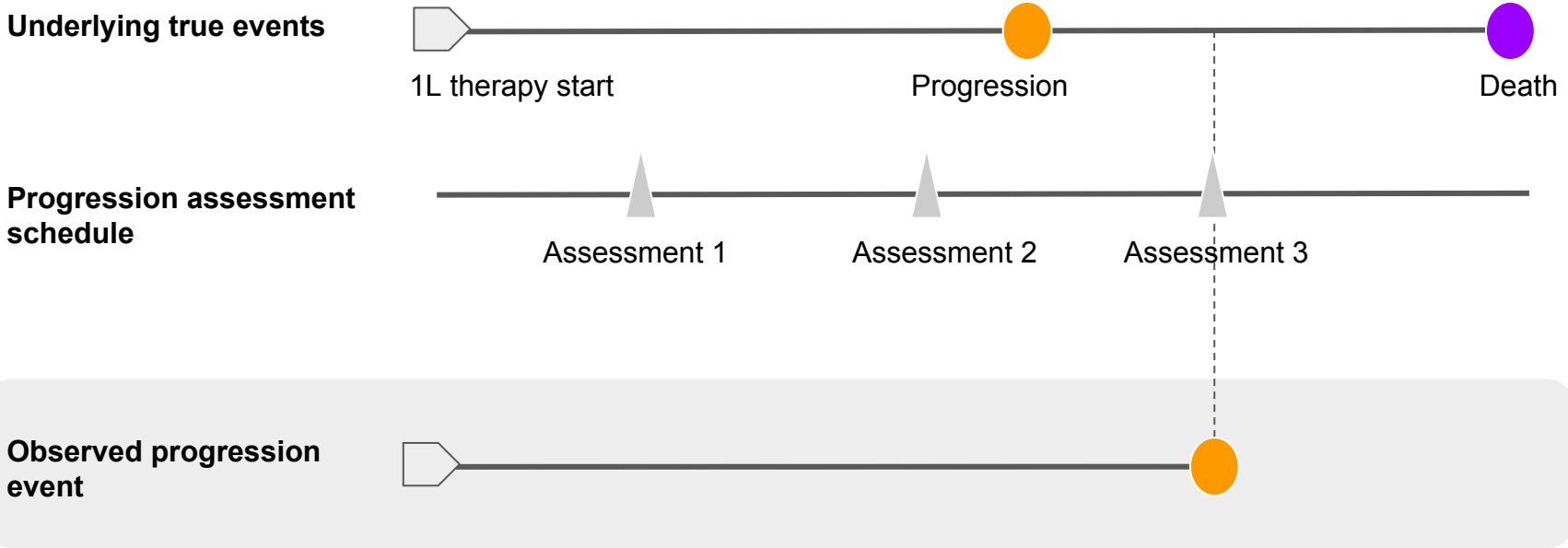
Underlying true events

Progression assessment done constantly

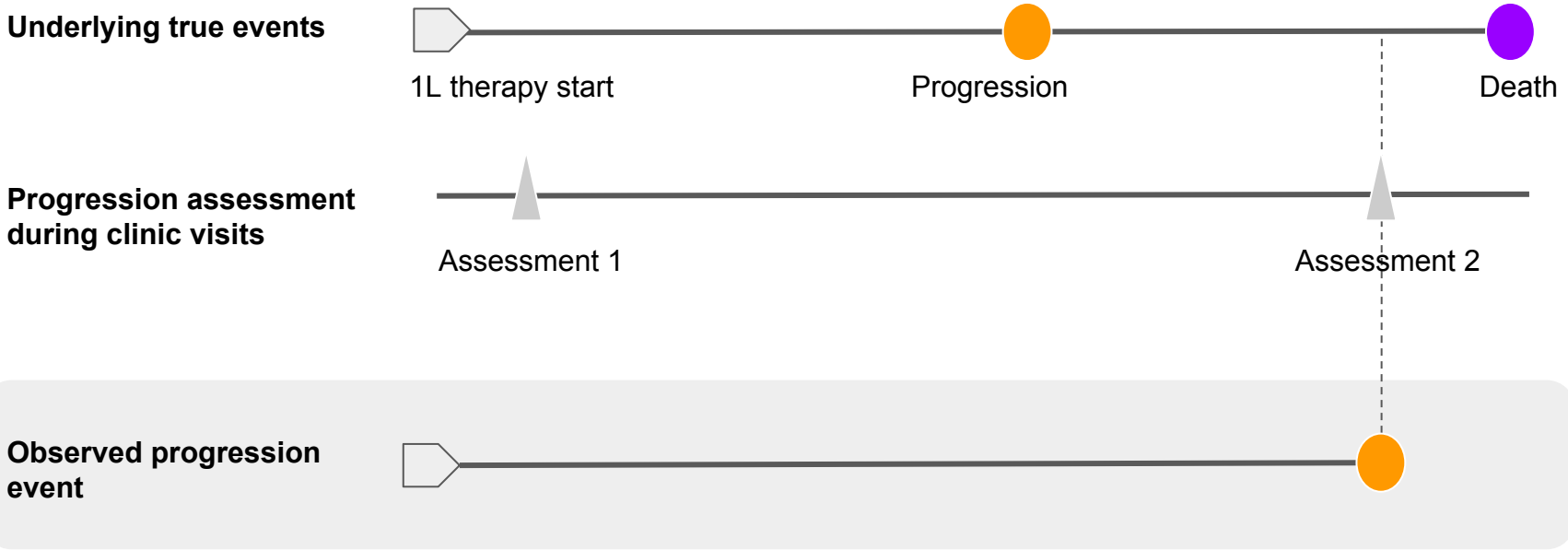
Observed progression event



Scan timing: Clinical Trial Setting



Scan timing: Real World Setting

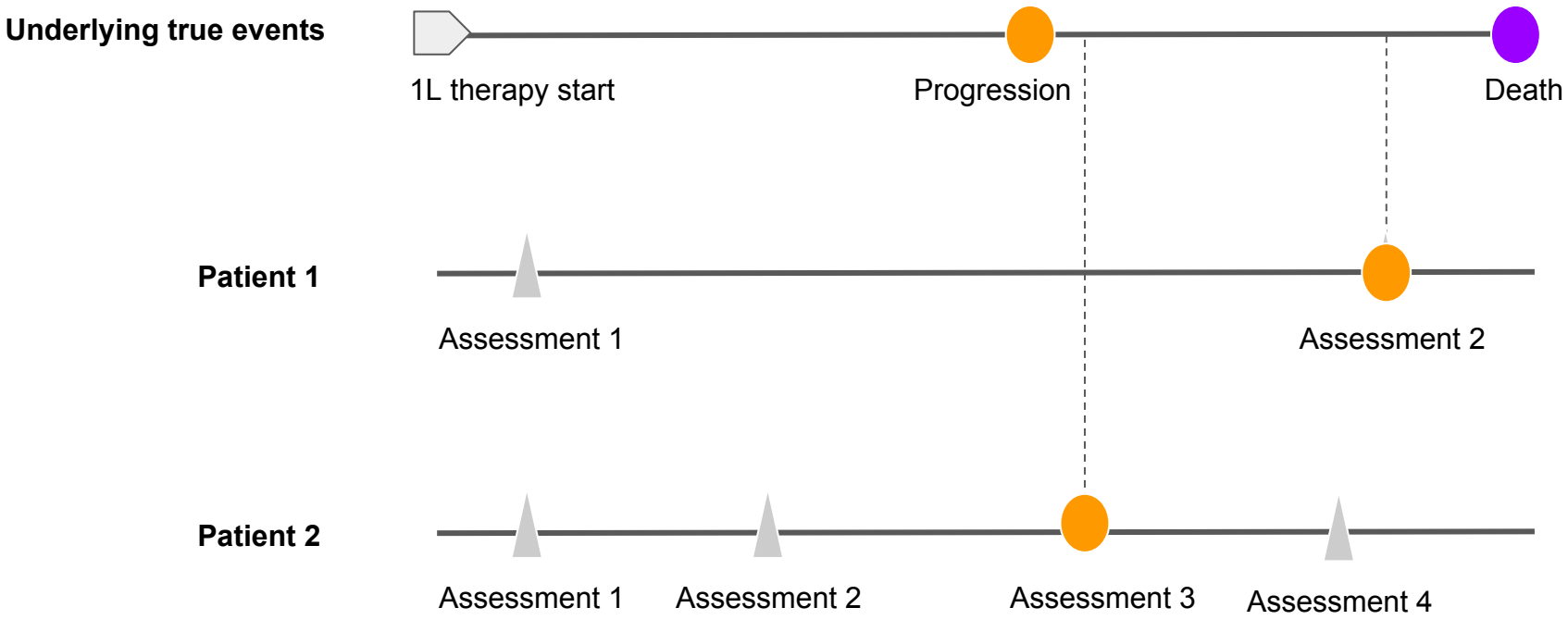


Surveillance Bias

Surveillance bias is where a patient's outcome appears better/worse, not because of more severe underlying disease, but rather because we have more/less opportunities to observe it leading to...

- biased estimates of progression-free survival (PFS) and other progression-based outcomes

- biased estimates of treatment effects when the assessment frequency differs between comparator groups



Goal

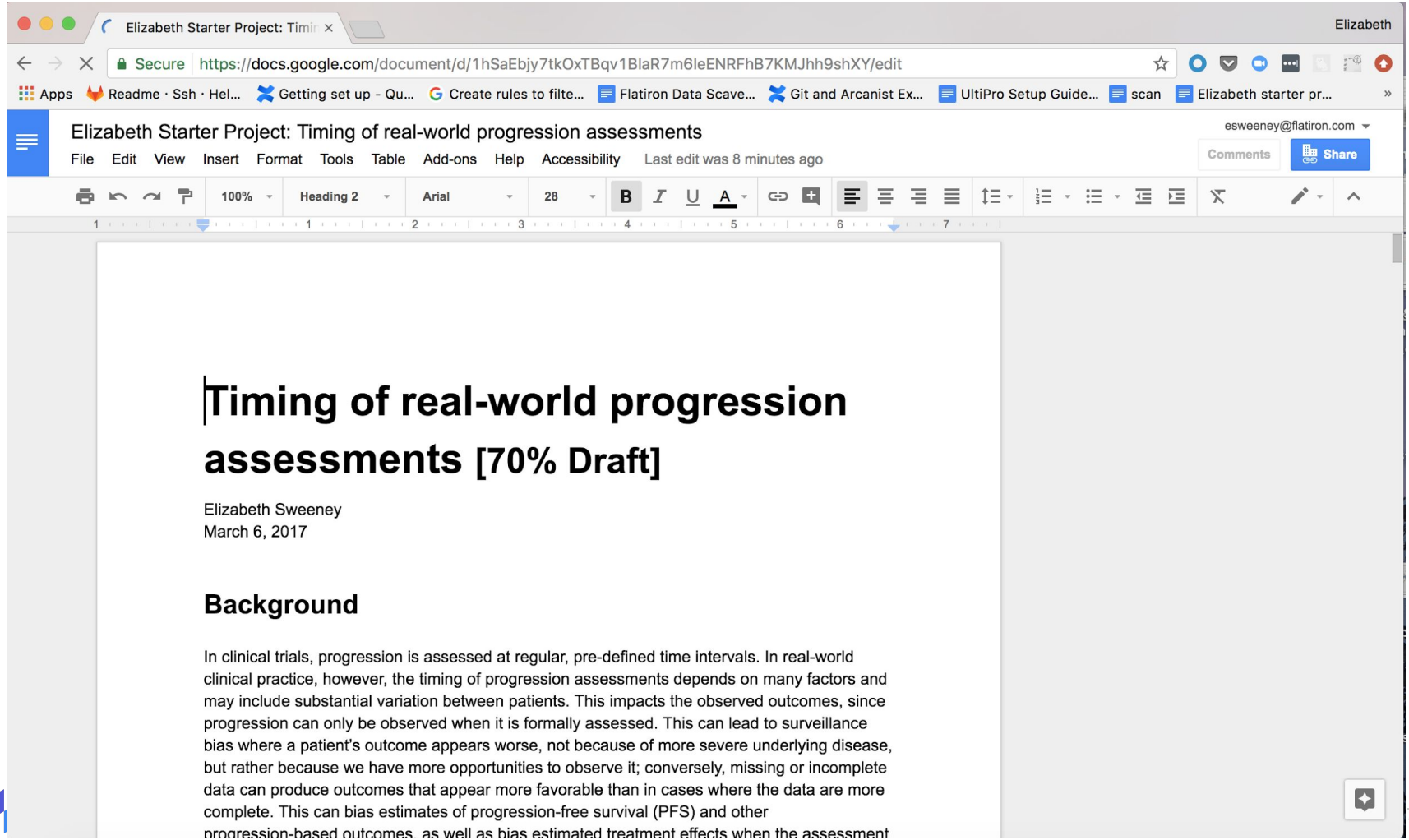
Determine the impact and biases introduced by differential progression assessment scan timing on real world progression in the aNSCLC patient cohort.

Flatiron Quantitative Scientist Workflow

1. Write Statistical Analysis Plan (SAP)
2. Feedback on SAP
3. Code
4. Code Review
5. Synthesize Results with the Team
6. Present final results to the Flatiron Research Oncology team
7. Publications

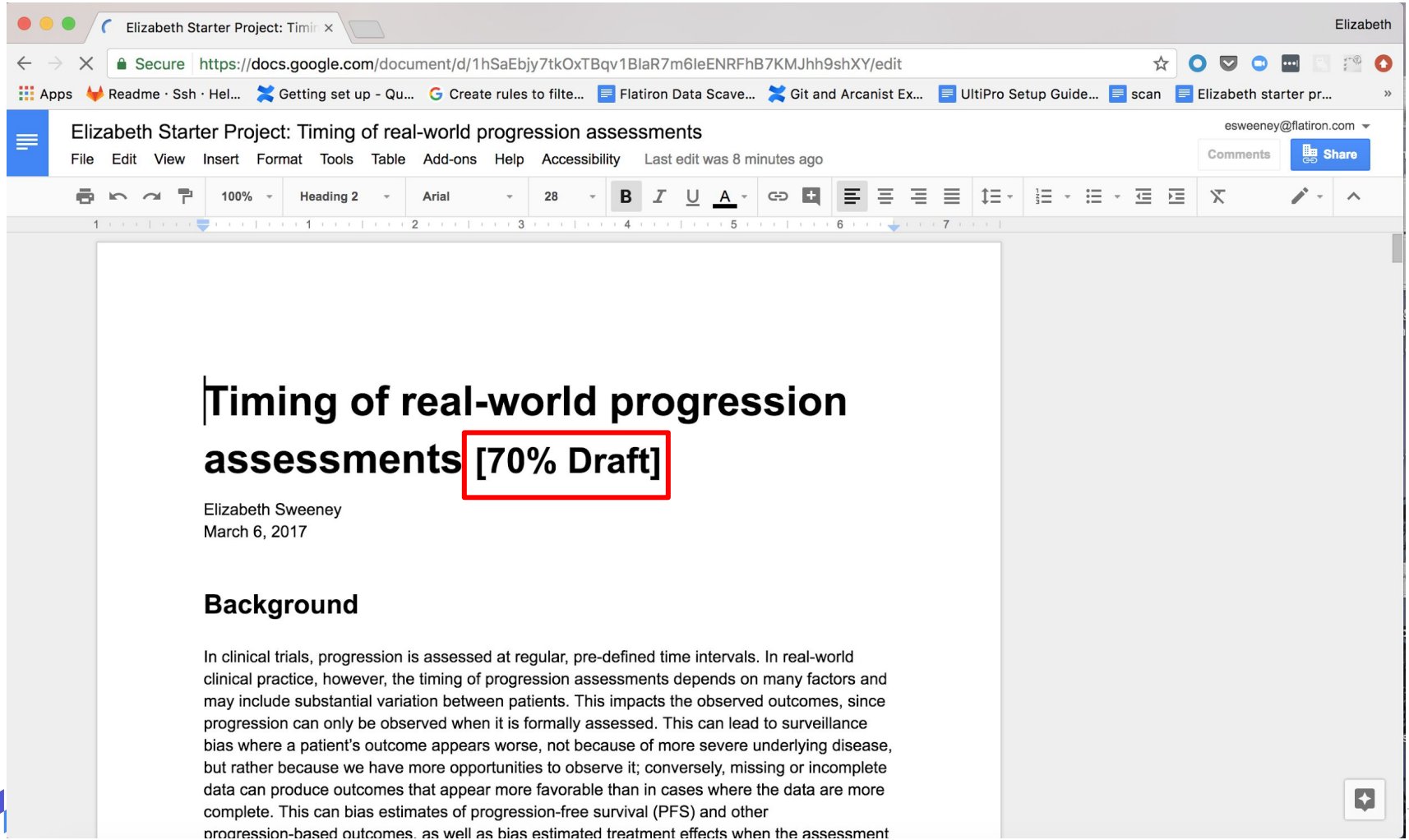
Flatiron Quantitative Scientist Workflow

1. Write Statistical Analysis Plan (SAP)
2. Feedback on SAP
3. Code
4. Code Review
5. Synthesize Results with the Team
6. Present final results to the Flatiron Research Oncology team


Elizabeth Starter Project: Timi
Elizabeth
Secure https://docs.google.com/document/d/1hSaEbjy7tkOxTBqv1BIaR7m6IeENRFhB7KMJhh9shXY/edit
Apps
Readme · Ssh · Hel...
Getting set up - Qu...
Create rules to filte...
Flatiron Data Scave...
Git and Arcanist Ex...
UltiPro Setup Guide...
scan
Elizabeth starter pr...
Elizabeth Starter Project: Timing of real-world progression assessments
esweeney@flatiron.com
File Edit View Insert Format Tools Table Add-ons Help Accessibility
Last edit was 8 minutes ago
Comments
Share
100%
Heading 2
Arial
28
Timing of real-world progression assessments [70% Draft]
Elizabeth Sweeney
March 6, 2017
Background
In clinical trials, progression is assessed at regular, pre-defined time intervals. In real-world clinical practice, however, the timing of progression assessments depends on many factors and may include substantial variation between patients. This impacts the observed outcomes, since progression can only be observed when it is formally assessed. This can lead to surveillance bias where a patient's outcome appears worse, not because of more severe underlying disease, but rather because we have more opportunities to observe it; conversely, missing or incomplete data can produce outcomes that appear more favorable than in cases where the data are more complete. This can bias estimates of progression-free survival (PFS) and other progression-based outcomes, as well as bias estimated treatment effects when the assessment

Flatiron Quantitative Scientist Workflow

1. Write Statistical Analysis Plan (SAP)
2. Feedback on SAP
3. Code
4. Code Review
5. Synthesize Results with the Team
6. Present final results to the Flatiron Research Oncology team


Elizabeth Starter Project: Timi
Elizabeth
Secure https://docs.google.com/document/d/1hSaEbjy7tkOxTBqv1BIaR7m6IeENRFhB7KMJhh9shXY/edit
Apps
Readme · Ssh · Hel...
Getting set up - Qu...
Create rules to filte...
Flatiron Data Scave...
Git and Arcanist Ex...
UltiPro Setup Guide...
scan
Elizabeth starter pr...
Elizabeth Starter Project: Timing of real-world progression assessments
esweeney@flatiron.com
File Edit View Insert Format Tools Table Add-ons Help Accessibility
Last edit was 8 minutes ago
Comments
Share
100%
Heading 2
Arial
28
Timing of real-world progression assessments [70% Draft]
Elizabeth Sweeney
March 6, 2017
Background
In clinical trials, progression is assessed at regular, pre-defined time intervals. In real-world clinical practice, however, the timing of progression assessments depends on many factors and may include substantial variation between patients. This impacts the observed outcomes, since progression can only be observed when it is formally assessed. This can lead to surveillance bias where a patient's outcome appears worse, not because of more severe underlying disease, but rather because we have more opportunities to observe it; conversely, missing or incomplete data can produce outcomes that appear more favorable than in cases where the data are more complete. This can bias estimates of progression-free survival (PFS) and other progression-based outcomes, as well as bias estimated treatment effects when the assessment

Flatiron Quantitative Scientist Workflow

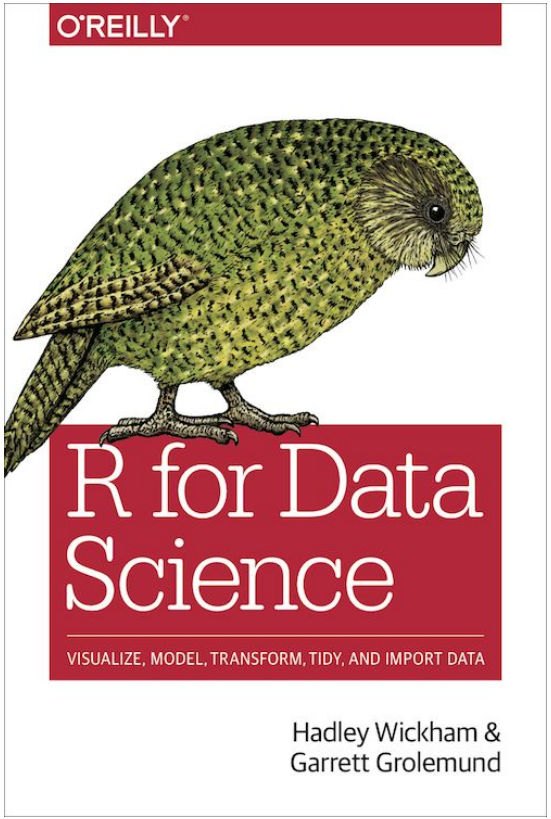
1. Write Statistical Analysis Plan (SAP)
2. Feedback on SAP
3. Code
4. Code Review
5. Synthesize Results with the Team
6. Present final results to the Flatiron Research Oncology team

- Use  and make R Markdown reports
- Internal R package called FlatironR
- R style guide that is modeled off of the [Google R Style guide]

```
GOOD:
ggplot(visit.summary, aes(ProSource, NumPatients, fill = Buckets)) +
  geom_bar(stat = "identity")

BAD:
ggplot(visit.summary, aes(ProSource, NumPatients, fill = Buckets)) +
geom_bar(stat = "identity")
```

- We use a lot of the Hadley Wickham packages (dplyr, purr, etc.)
- Reading group for R for Data Science
- Bi-weekly R working group



Flatiron Quantitative Scientist Workflow

1. Write Statistical Analysis Plan (SAP)
2. Feedback on SAP
3. Code
4. Code Review
5. Synthesize Results with the Team
6. Present final results to the Flatiron Research Oncology team

- Code review using Phabricator
- Phabricator is a suite of web-based software development collaboration tools, including the *Differential* code review tool

Diff 63616

research/esweeney/Starter_Project/Elizabeth_Starter_Project.Rmd

View Options

This file was added.

```
---
title: "Timing of real-world progression assessments"
output:
  html_document:
    toc: true
    toc_float: true
---

```{r global_options, include=FALSE}
knitr::opts_chunk$set(fig.width  = 8,
                      fig.height = 5,
                      echo       = FALSE,
                      warning    = FALSE,
                      message    = FALSE,
                      tidy       = F)

# Disable scientific notation
options(scipen = 999)

```
```

This file was added.

```
---
title: "Timing of real-world progression assessments"
output:
  html_document:
    toc: true
    toc_float: true
---

```{r global_options, include=FALSE}
knitr::opts_chunk$set(fig.width  = 8,
                      fig.height = 5,
                      echo       = FALSE,
                      warning    = FALSE,
```

New Inline Comment

Change warning to = TRUE

Cancel | Save Draft
```

```
#####################################
## Source Functions
#####################################
source('~/code/flatiron/research/esweeney/Starter_Project/rwp_nsclc_edm_functi
ons.R')

#####################################
## Packages
#####################################
LoadPackages(c('FlatironR',
               'broom',
               'dplyr',
               'ggplot2',
               'gridExtra',
               'lubridate',
               'irr',
               'purrr',
               'scales',
               'survival',
               'SurvCorr',
               'tidyr',
               'DT',
               'mada',
               'compareGroups',
               'knitr',
               'rmarkdown'))
```

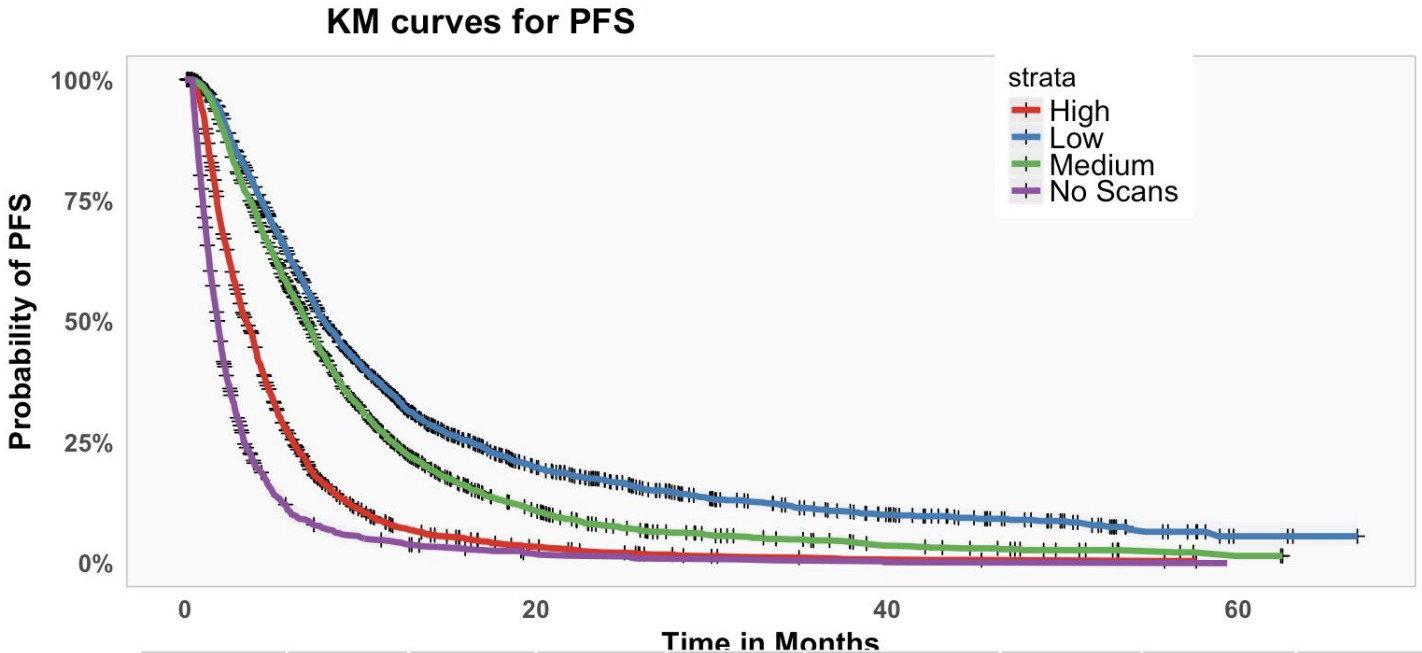
```
34 #####################################
35 ## Source Functions
36 #####################################
37 source('~/code/flatiron/research/esweeney/Starter_Project/Code_from_Paul/rwp_ns
   clc_edm_functions.R')
38
39 #####################################
40 ## Packages
41 #####################################
42 LoadPackages(c('FlatironR',
43                'broom',
44                'dplyr',
45                'ggplot2',
46                'gridExtra',
47                'lubridate',
48                'irr',
49                'purrr',
50                'scales',
51                'survival',
52                'SurvCorr',
53                'tidyr',
54                'DT',
55                'mada',
56                'compareGroups',
57                'knitr',
58                'rmarkdown',
59                'survminer'))
60
```

Flatiron Quantitative Scientist Workflow

1. Write Statistical Analysis Plan (SAP)
2. Feedback on SAP
3. Code
4. Code Review
5. Synthesize Results with the Team
6. Present final results to the Flatiron Research Oncology team

Goal

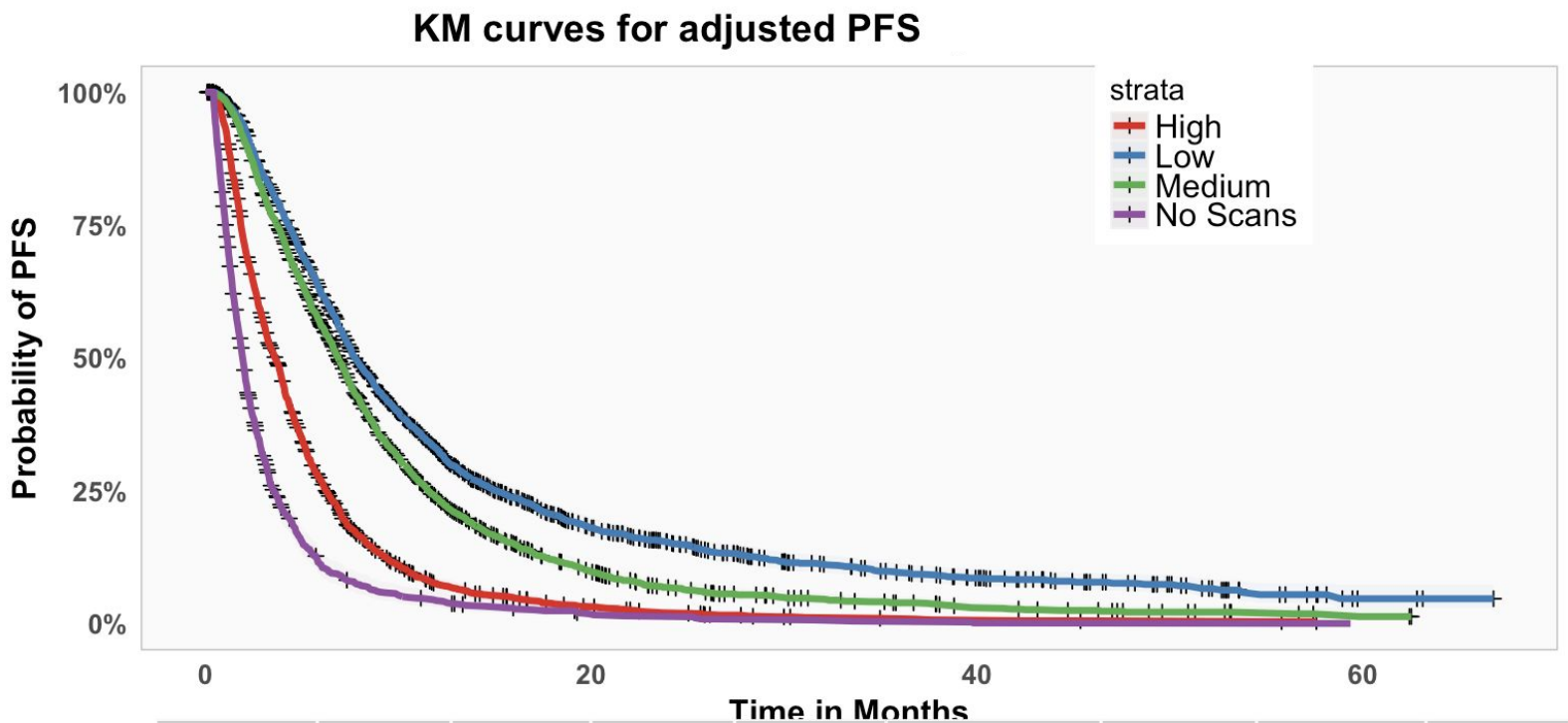
Determine the impact and biases introduced by differential progression assessment scan timing on real world progression in the aNSCLC patient cohort.



	records	n.max	n.start	events	*rmean	*se(rmean)	median	0.95LCL	0.95UCL
High	2560	2560	2560	2168	5.38	0.16	3.54	3.28	3.77
Low	2461	2461	2461	1813	14.41	0.4	7.97	7.61	8.46
Medium	2353	2353	2353	1950	10.3	0.27	6.92	6.66	7.25
No Scans	1137	1137	1137	873	3.34	0.19	1.9	1.77	2

Survival curves adjusted for potential confounders that could be causing the differences between the groups:

1. age at advanced diagnosis
2. gender
3. region
4. race
5. diagnosis stage
6. biomarker status



	records	n.max	n.start	events	*rmean	*se(rmean)	median	0.95LCL	0.95UCL
Low	2461	2461	2461	1813	13.63	0.35	7.77	7.51	8.2
Medium	2353	2353	2353	1950	9.93	0.24	6.92	6.66	7.21
High	2560	2560	2560	2168	5.38	0.16	3.64	3.41	3.87
No Scans	1137	1137	1137	873	3.41	0.19	1.97	1.84	2.07

Practical Implications

- Estimates of PFS can be biased depending on how frequently patients are scanned
- This can impact our ability to estimate treatment effects, especially if the scanning frequency of patients differs between a treatment and a control group

Flatiron Quantitative Scientist Workflow

1. Write Statistical Analysis Plan (SAP)
2. Feedback on SAP
3. Code
4. Code Review
5. Synthesize Results with the Team
6. Present final results to the Flatiron Research Oncology team

Research Oncology Collaboration

Weekly meetings with the research oncology team where we present the output of our analysis to get feedback and direction

Often share output in the form of R Markdown documents

My Workflow Differences

Flatiron

1. Write Statistical Analysis Plan (SAP)
2. Feedback on SAP
3. Code
4. Code Review
5. Synthesize Results with the Team
6. Present final results to the Flatiron Research Oncology team
7. Publication

Academic

1. Research Question
2. Code
3. Write Publication

Questions?