Understanding and reacting to the impact of COVID-19 by combining statistics with visualization

Basel Biometrics Section Seminar: Graphics for decision-making in biomedical research and drug development

March 8, 2021
Topics

• Flatiron Health and real-world evidence

• A principle we’ll draw upon in using visualization to aid quantitative analysis: generate hypotheses

• Application: understanding the impact of COVID-19 on Flatiron Health data
Real-world data and real-world evidence

• **RW data** are *data* relating to patients health status and delivery of healthcare routinely collected from various sources.

• **Real-world evidence (RWE)** is derived from real-world data (RWD) through the **application of research methods**.
Flatiron’s approach: developing high-quality real world evidence

“RWE is evidence derived from RWD through the application of research methods”
RWD curation: structured and unstructured data

- Demographics
- Diagnosis
- Visits
- Labs
- Therapies
- Pathology
- Discharge Notes
- Physician Notes
- Radiology & Biomarker Reports

Structured Data Processing
Unstructured Data Processing

Research Database

© Flatiron Health
Topics

- Flatiron Health and real-world evidence
- A principle we’ll draw upon in using visualization to aid quantitative analysis: generate hypotheses
- Application: understanding the impact of COVID-19 on Flatiron Health data
A question we’ll hear: if we’d like to understand whether two variables are correlated... why use visualization? Can we not simply compute the correlation statistic?

• Imagine a ‘vehicle’ database table* with the following variables:
  • Economic performance (mpg)
  • Cylinders
  • Displacement
  • Power (hp)
  • Weight (lbs.)
  • 0-60 mph (s)
  • Year

• An analytical question may be: is the power (hp) of a vehicle correlated with its economic performance (mpg)?

• Yet just one analysis here may lead to deeper questions: which dimensions are correlated, and under what conditions?
  • We could prepare a correlation statistic on each possible pair of variables...
    • Yet... what if a given correlation only emerges among vehicles designed prior to 1990...
      • That weight less than 4000 lbs...?
    • And what if we’re not certain of the pertinent subsets before we begin?
      • Strictly-computational approaches exist that will search for high-performing model specifications...
      • Yet... what if we have outside knowledge that should be incorporated into this choice of a specification?
      • Some computational approaches will also allow us to incorporate hypotheses and other tacit information (e.g., by beginning the search in the a-priori believed most-promising areas)
      • Interactive visualization allows us to formulate and update these hypotheses in real time, potentially avoiding unhelpful – and perhaps even infeasible – computation


Example on next slide

*In this example, we’ll utilize the ‘Motor Trend Car Road Tests’ database (built into the R programming language)

© 2014 Neil McQuarrie
Example

Topics

• Flatiron Health and real-world evidence

• A principle we’ll draw upon in using visualization to aid quantitative analysis: generate hypotheses

• Application: understanding the impact of COVID-19 on Flatiron Health data
In April and May of 2020, we noticed that the survival of patients in our advanced NSCLC datamart – under certain ways of measuring this survival – began to break trend.

Is NSCLC patients’ survival affected by COVID-19?

What is happening?
<table>
<thead>
<tr>
<th>Hypothesis sequence</th>
<th>Visualization-driven finding</th>
<th>Decision</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. A COVID-driven increase in mortality was reducing overall survival among NSCLC patients</td>
<td>Slightly <em>fewer</em> mortality events per NSCLC patient were actually present in our data than over the same time period one year prior</td>
<td>Explore patterns in any additional variables that can contribute to overall survival calculations</td>
</tr>
<tr>
<td>2. Something unusual is happening with the data that one may use for longitudinal censoring</td>
<td>Patients’ most-recently recorded clinic visit dates were in fact skewing earlier</td>
<td></td>
</tr>
</tbody>
</table>
Our first hypothesis was that a COVID-driven increase in mortality was reducing overall survival. **A visualization of the accrual of mortality events among samples of patients from our May 2020 versus May 2019 datamarts showed that this was not actually as imaginable**

Cumulative # of mortality events, among random samples of NSCLC patients diagnosed prior to 2020 and prior to 2019, by week of the year

- **March 7** (In 2020, NYC declares state of emergency)
- **36 fewer YTD mortality events had accrued in the May 2020 vs May 2019 datamart delivery**

Included = patients in the NSCLC datamart with an advanced NSCLC diagnosis prior to the time period under observation
Second, we looked at the distribution of patients’ *most-recently recorded clinic visit dates* which, under certain approaches to calculating overall survival, play a role in longitudinal censoring.

The May 2020 datamart’s distribution of patients’ most-recently recorded clinic visit dates was skewed toward times prior to NYC declaring its state of emergency.

Certain approaches to calculating overall survival will utilize a composite of a patient’s most-recently recorded clinic visit date and date of death (if one exists) to censor patients from that calculation.

Under such an approach, a slowing in the accrual of final visit records could have led to a small, but noticeable – and yet artificial – decrease in calculated overall survival.
In conclusion

• When calculated using Flatiron’s May 2020 datamart, certain survival-related statistics were showing a small but noticeable downward break from the prior trend

• Exploratory visualization helped us see that our first hypothesis – that, overall, patients at our clinics were passing away sooner, conceivably as a result of COVID-19 – was less imaginable than first believed

• Instead, we saw that changes in patient visit patterns were more likely to be driving these trends in measured statistics

• From here, we were able to make certain that any further statistical analysis was performed with this dynamic in mind