OHDSI community efforts on COVID-19 disease natural history: Status update and look forward to ‘life after COVID’

Patrick Ryan, PhD
Janssen Research and Development
Columbia University Irving Medical Center

on behalf of OHDSI community:
CHARYBDIS study leads: Anthony Sena, Kristin Kostka, Talita Duarte-Salles, Albert Prats-Uribe
Open collaboration requires FULL transparency in every step of the research process

• Protocol and analysis source code freely available and directly downloadable: https://github.com/ohdsi-studies/Covid19CharacterizationCharybdis

• Phenotype definitions are both human-readable and computer-executable using ATLAS against any OMOP CDM: https://atlas.ohdsi.org/

• All analysis results will be available for public exploration through interactive R shiny application: http://data.ohdsi.org/Covid19CharacterizationCHARYBDIS/

• The study is a living evidence repository: any data partners can execute analysis and share aggregate results at any point, including updates as data accumulate

Join the Journey!
CHARYBDIS target cohorts

- Persons tested for SARS-COV-2
  - Persons tested positive for SARS-COV-2
    - Persons hospitalized with positive test for SARS-COV-2
      - Persons hospitalized and requiring intensive services with positive test for SARS-COV-2
        - POTENTIAL ADDITION?: Persons post-discharge after COVID-19 diagnosis record OR a SARS-COV-2 positive test
    - Persons hospitalized with COVID-19 diagnosis record OR a SARS-COV-2 positive test
  - Persons with COVID-19 diagnosis record OR a SARS-COV-2 positive test
    - Persons hospitalized and requiring intensive services with COVID-19 diagnosis record OR a SARS-COV-2 positive test
CHARYBDIS subgroup cohorts

Persons tested for SARS-COV-2

Persons tested positive for SARS-COV-2

Persons with positive test for SARS-COV-2

Persons hospitalized with positive test for SARS-COV-2

Persons hospitalized and requiring intensive services with positive test for SARS-COV-2

Persons with COVID-19 diagnosis record OR a SARS-COV-2 positive test

Persons hospitalized with COVID-19 diagnosis record OR a SARS-COV-2 positive test

Persons hospitalized and requiring intensive services with COVID-19 diagnosis record OR a SARS-COV-2 positive test

Stratification cohorts:
- Age: <18, >65
- Gender: Female/Male
- Race: Black/White
- Index month
- Hypertension
- Type 2 Diabetes
- Heart disease
- Obesity
- Asthma
- COPD
- Chronic kidney disease
- End stage renal disease
- Cancer
- Autoimmune conditions
- Dementia
- HIV
- Pregnant women
- Follow-up time: >=30d
CHARYBDIS Time windows

Cohort start date = Index date

-365d to -1d

-30d to -1d

0d

0d to 30d

0d to 365d

31d to 365d

Pre-index characteristics for medical history:
Demographics:
- Age group (5-year strata)
- Sex
Concept-based:
- Condition groups (SNOMED + descendants), >=1 occurrence during the interval
- Drug era groups (ATC/RxNorm + descendants), >=1 day during the interval which overlaps with at least 1 drug era

Cohort features:
- Symptoms (fever, cough, malaise, myalgia, dyspnea)
- Acute clinical events (AKI, ARDS, AMI, PE/DVT, …)
- Service utilization (hospitalization, ventilation, tracheostomy, ECMO, dialysis)

Post-index characteristics for treatments and outcomes:
Concept-based:
- Condition groups (SNOMED + descendants), >=1 occurrence during the interval
- Drug era groups (ATC/RxNorm + descendants), >=1 day during the interval which overlaps with at least 1 drug era

Cohort features:
- Symptoms (fever, cough, malaise, myalgia, dyspnea)
- Acute clinical events (AKI, ARDS, AMI, PE/DVT, …)
- Service utilization (hospitalization, ventilation, tracheostomy, ECMO, dialysis)

Potential additions?
## Data partners contributing to CHARYBDIS thusfar

<table>
<thead>
<tr>
<th>Database name</th>
<th>Geography</th>
<th>Data type</th>
</tr>
</thead>
<tbody>
<tr>
<td>Premier</td>
<td>US (National)</td>
<td>Hospital billing</td>
</tr>
<tr>
<td>Optum EHR</td>
<td>US (National)</td>
<td>Electronic health records</td>
</tr>
<tr>
<td>Iqvia Open Claims</td>
<td>US (National)</td>
<td>Administrative claims</td>
</tr>
<tr>
<td>VINCI (VA)</td>
<td>US (National)</td>
<td>Electronic health records</td>
</tr>
<tr>
<td>STARR (Stanford)</td>
<td>US (CA)</td>
<td>Electronic health records</td>
</tr>
<tr>
<td>TRDW (Tufts)</td>
<td>US (MA)</td>
<td>Electronic health records</td>
</tr>
<tr>
<td>CUIMC (Columbia)</td>
<td>US (NY)</td>
<td>Electronic health records</td>
</tr>
<tr>
<td>SIDIAP</td>
<td>Spain</td>
<td>Electronic health records</td>
</tr>
<tr>
<td>SIDIAP-H</td>
<td>Spain</td>
<td>EHR-hospital linkage</td>
</tr>
<tr>
<td>HM Hospitales</td>
<td>Spain</td>
<td>Hospital billing</td>
</tr>
<tr>
<td>ICPI</td>
<td>Netherlands</td>
<td>Electronic health records</td>
</tr>
<tr>
<td>CPRD</td>
<td>UK</td>
<td>Electronic health records</td>
</tr>
<tr>
<td>HIRA</td>
<td>South Korea</td>
<td>Administrative claims</td>
</tr>
<tr>
<td>DCMC</td>
<td>South Korea</td>
<td>Electronic health records</td>
</tr>
</tbody>
</table>

All databases standardized to OMOP CDM v5.3

---

Diagram: Standardization processes and data elements.
Live demo of CHARYBDIS

https://data.ohdsi.org/Covid19CharacterizationCharybdis/
Live demo of CHARYBDIS

https://data.ohdsi.org/Covid19CharacterizationCharybdis/
Live demo of CHARYBDIS

[Image of a screen displaying a live demo of CHARYBDIS, showing a table and a graph with data related to COVID-19 and SARS-CoV-2]

https://data.ohdsi.org/Covid19CharacterizationCharybdis/
Live demo of CHARYBDIS

https://data.ohdsi.org/Covid19CharacterizationCharybdis/
Live demo of CHARYBDIS cohort diagnostics

https://data.ohdsi.org/Covid19CharacterizationCharybdisDiagStrata/
Live demo of CHARYBDIS cohort diagnostics

https://data.ohdsi.org/Covid19CharacterizationCharybdisDiagStrata/
Using Twitter to characterize the COVID disease natural history and 'life after COVID'

Juan M. Banda
www.panacealab.org
Georgia State University
Preface: Twitter is gaining attention for health-related research since 2009

![Number of publications per year](image)

Results of PubMed Query for Twitter and Health
Benefits of using Twitter:

1) Good population representation
2) Everybody can post and have an account
3) Anonymity = unfiltered opinions
4) Data is freely available*
5) Tons of data generated each day (hundreds of millions of tweets get posted every day)
6) Easy filtering (hashtag usage, people mentions)

Traditional disadvantages of using Twitter:

• Messy data (plenty of misspellings, shorthand, emojis, etc.)
  • There are at least 25 different ways people misspell hydroxychloroquine

• Attribution is an issue – are people just mentioning something or did it happen to them?

• Freely available data is only a 1% sample of whole set

• Collection is hard and needs to be ongoing for days/weeks before getting considerable mass
The COVID opportunity – highly focused data

The dataset:

• 490+ Million Tweets

• ONLY COVID related chatter is included

Longitudinal – January 27th to today… and growing

Dataset: https://doi.org/10.5281/zenodo.3723939
Recent additions: https://github.com/thepanacealab/covid19_twitter
Current work: Drug characterization

• Methods to deal with misspellings and noisiness of data:

• Charybdis-like characterization over countries (work with Dani Prieto-Alhambra – University of Oxford)

Table 2. Drug ingredient mentions found

<table>
<thead>
<tr>
<th>Drug Ingredient</th>
<th>Frequency</th>
</tr>
</thead>
<tbody>
<tr>
<td>hydroxychloroquine</td>
<td>204,879</td>
</tr>
<tr>
<td>remdesivir</td>
<td>72,841</td>
</tr>
<tr>
<td>chloroquine</td>
<td>49,915</td>
</tr>
<tr>
<td>oxygen</td>
<td>37,961</td>
</tr>
<tr>
<td>vitamin D</td>
<td>25,445</td>
</tr>
<tr>
<td>dexamethasone</td>
<td>25,142</td>
</tr>
<tr>
<td>zinc</td>
<td>24,843</td>
</tr>
<tr>
<td>azithromycin</td>
<td>16,079</td>
</tr>
<tr>
<td>ibuprofen</td>
<td>8,469</td>
</tr>
<tr>
<td>ivermectin</td>
<td>6,390</td>
</tr>
</tbody>
</table>

Figure 1. Timeline of Tweets with potential drug treatment mentions.

* https://openreview.net/forum?id=qIgPXs9FWa
Current Work: Symptom/condition detection

• Self-reported symptoms on Twitter vs EHR lists *
  • Can we find related symptoms both found on EHR’s (Callahan, A., Steinberg, E., Fries, J.A. et al. Estimating the efficacy of symptom-based screening for COVID-19. npj Digit. Med. 3, 95 (2020). https://doi.org/10.1038/s41746-020-0300-0) but on Twitter?

<table>
<thead>
<tr>
<th>Term</th>
<th>Frequency</th>
</tr>
</thead>
<tbody>
<tr>
<td>pneumonia</td>
<td>110124</td>
</tr>
<tr>
<td>infection</td>
<td>71882</td>
</tr>
<tr>
<td>influenza</td>
<td>36390</td>
</tr>
<tr>
<td>cough</td>
<td>35753</td>
</tr>
<tr>
<td>anxiety</td>
<td>34658</td>
</tr>
<tr>
<td>pain</td>
<td>12773</td>
</tr>
<tr>
<td>depression</td>
<td>12189</td>
</tr>
<tr>
<td>asthma</td>
<td>8307</td>
</tr>
</tbody>
</table>

* https://github.com/thepanacealab/covid19_biohackathon/tree/master/user_symptoms
What does this lead to?

• Since we can find symptoms and drugs, we can also find people that had COVID and their symptoms after infection!

• On-going work with Dani Prieto-Alhambra and others
  • Incorporates methods shown before + manual review by clinicians

Some very preliminary findings:

<table>
<thead>
<tr>
<th>Symptom</th>
<th>Patients with symptom, %</th>
</tr>
</thead>
<tbody>
<tr>
<td>Fatigue</td>
<td>789</td>
</tr>
<tr>
<td>Shortness of breath−dyspnea</td>
<td>701</td>
</tr>
<tr>
<td>Chest pain</td>
<td>687</td>
</tr>
<tr>
<td>Palpitations</td>
<td>674</td>
</tr>
<tr>
<td>Anxiety</td>
<td>212</td>
</tr>
<tr>
<td>Post-exertional malaise</td>
<td>36</td>
</tr>
<tr>
<td>Tired = fatigue</td>
<td>36</td>
</tr>
<tr>
<td>Muscle pain = myalgia</td>
<td>35</td>
</tr>
</tbody>
</table>

The gory details:

• Technical stuff:
  • “Building tools and frameworks for large-scale social media mining: Creating data infrastructure for COVID-19 research” dair.ai meetup 7/22: https://www.meetup.com/dair-ai/events/271690722/

• Extended version of today’s short talk: