Forecasting Patient Enrolment for Clinical Trials

Team Project Sprint Review
April 3rd, 2020

01

BW-Cloud

Why and how we choose BW-Cloud as our provider

02

Clinical Trials Data

What we already achieved using the clinical trials API

03

Data Enrichment

Additional sources to enrich the clinical trials data

04

Next Steps

Plans on how to proceed in this project

BW-Cloud

- What is BW-cloud? What is used for?
- Its configuration
- Its limitation

BW-Cloud

Bw-Cloud is an "infrastructure-as-a-service" environment, specially developed and operated for research and teaching in Baden-Württemberg[1].

Its free offer	Our initial demand
• Disk space: 12GB • Ram: 1GB • OS: Ubuntu 18.04	• 2GB of data from clinicaltrials.gov • Simple, lightweight database • Python3 and some libraries

=> Should be enough for initial steps

[1]: https://www.bw-cloud.org/en/first_steps

BW-Cloud

The limitations

- **Backup**: No provided automatic backup option[1]. Currently have to do the backup manually. The snapshot images will be stored here.

- **Hardware limitation**: The current configuration is only enough for initial steps. May need better resource in the future

1: https://www.bw-cloud.org/en/faq/backup

Clinical Trials Data

- Which query types does the API of clinicaltrials.gov offer?
- Our approach to extract all relevant studies
- The implementation in python
- Which challenges did we encounter?

ClinicalTrials.gov API – Query Types

Query URL Type	Output	Restriction
Full Studies	all content from the first study record returned	100 study records
Study Fields	the values of one or more fields from up to all study records returned	1,000 study records
Field Values	unique list of values for one study field from all study records returned	Unlimited

ClinicalTrials.gov API – Approach

1. Get a list of trials IDs (NCTId)

- Relevant studies only: completed and in phase 3 or 4
- Returns 33720 records

2. Get specific information about every trial

- Study Fields: up to 20 attributes (first attempt)
- Full studies: all attributes (next step)

ClinicalTrials.gov API – Code

ClinicalTrials.gov API – Challenges

Data amount for each study

- Problem: specific queries vs. complete queries (regarding number of attributes)
- Solution: retrieve as much data as possible

Different data structure

- Problem: different handling needed, depending if the attribute is a list or single value
- Solution: analyze data and identify possible returns of every attribute

Special characters in the data

- Problem: SQL Query
- Solution: replace critical characters (problem: data manipulation)

Data Enrichment

- Find additional data sources
- Combine with the main data source clinicaltrials.gov
- Goal: Make more precise forecasts

Possible Sources

Hospital Data

Attributes focus on research activities

No API

The website has only 822 German Hospitals included, but there are 1942 in Germany

... maybe we need to find country specific hospital data for each country, because they have a much better quality

Corruption Data

Transparency international offers information on corruption in 180 countries

Easily accessible as .xlsx

Also provides each countries world rank and the standard deviation of the corruption index

Diseases Burden

Cause of death given democratic(sex, age), time(daily, year), and geography(global, regional, country-level) factors
In year 2000-2016

Easily accessible as .xls

...has different categories of the cause of death, will need to do categorization to integrate the dataset

Possible Sources

Country Attributes

Organization for economic co-operation and development

Worldbank databank

→ Export as CSV or XML files

Wikidata API
→ Retrieve data via HTTP requests

World Population

Population density of each country

United Nation dataset
→City population with a lot of features. API provided

World Bank dataset
→ CSV, XML, XLSX are available; no API provided

Legal Factors

Has 13 countries data of regulations with quick factors which includes:

Application language,
Age of minors,
Registry fee,
Ethics Committee Fees...

--- has only 13 countries, but it is the most concrete one so far

Next Steps

Next Steps

- Code improvements
  - Retrieve full studies instead of only 20 attributes
  - Separation of functions to make it clearer
- Find further API or datasets
- Integrate data into our basis dataset
- Data Preprocessing
  - Feature Selection, Encoding, Data Analysis

Appendix

Country Attribute Data

- Organization for economic co-operation and development
- https://stats.oecd.org/Index.aspx?DataSetCode=REGION_DEMOGR
- Regional and country data
  - 2 000 regions in 36 OECD countries
  - Yearly time series data
  - 40 indicators of demography, economic accounts, labour market, social and innovation themes in the OECD member countries and other economies.
  - Also Social and environmental factors like health access
- Export as Excel, csv or xml file

Country Demographics Data

- **Worldbank databank**
- https://databank.worldbank.org/source/health-nutrition-and-population-statistics#
- Possible data: different age population, birth rate, children living with HIV, health data in country
- Lot of possible attributes but not all data available
- Extract as csv, excel and tabbed txt
- No api available

- **Wikidata API**
- Possible data formats = json, xml, rdf
- REST API: https://de.wikipedia.org/wiki/Wikipedia:Technik/Datenbank/API#REST

Hospital Data

https://hospitals.webometrics.info/en/Americas/USA

Attributes:

- World Rank (combined values of size, visibility, rich files and scholar)
- Hospital Name (-> City)
- Country
- Size (by page rank of search engines)
- Visibility (total number of unique external links received)
- Rich Files (evaluation of their relevance to academic and publication activities)
- Scholar (number of papers and citations for each academic domain)
-

-> only 822 German Hospitals included, but there are 1942 in Germany and only 3619 US hosptials, but there are 6,146

Additional Links

Number of specialized hospitals per 10 000 inhabitants: http://apps.who.int/gho/data/view.main.30000

German Hospitals by state (number of beds, patients, occupancy...): https://www.destatis.de/DE/Themen/Gesellschaft-Umwelt/Gesundheit/Krankenhaeuser/Tabellen

All German hospitals by region, city, treated diseases and quality: https://www.deutsches-krankenhaus-verzeichnis.de

More detailed data on US hospitals by state, beds, kind of hosptial: https://www.aha.org/statistics/fast-facts-us-hospitals

Corruption Data

Link from the sovanta slides: https://www.transparency.de/cpi/
Table view: https://www.transparency.de/cpi/cpi-2019/cpi-2019-tabellarische-rangliste/#

You can easily download the whole data set as .xlsx at the website of transparency international:
https://www.transparency.org/cpi2019

Attributes:

- **Country (and ISO Country Code)**
- **Corruption Index**
- **Rank**
- **Standard error and confidence intervals**
- **Number of Sources**

Maybe the degree of corruption can be an indicator how long the clinical trial is going to take or how much it is going to cost

Legal Factors

ClinRegs

- 13 countries legal related data
  - has 13 countries data of regulations(Brazil, China, India, Kenya, Liberia, Malawi, Peru, South Africa, Tanzania, Thailand, Uganda, the United Kingdom, and the United States. Mali, Mexico, Haiti, and Vietnam) and quick factors for comparison

WHO international clinical trial regulations

- The WHO Registry Criteria have been categorized into six main areas: Content Quality and Validity Accessibility Unambiguous Identification Technical Capacity Administration and Governance

Population density data

- World bank dataset: CSV, XML, XLSX are available, containing data until 2018 by country
- UN Data: numbers of population for each city, has a lot of additional features.
  Document for API is also provided here