Synthetic Patient Appointment Dataset

1. Dataset description

A synthetic dataset of patient appointments, referrals, and journeys to a fictional service in the North East of England. The code can be adjusted to incorporate any area on mainland Great Britain. NI or the islands can be integrated too, however the structure of postcode, GP and OSA public data is different, and data input handlers will need to be adjusted.

The behaviour of the patients (visiting their nearby GP followed by attending a specialist clinic), appointments (clinic appointments within 7day-6weeks of the referral (gp appointment)), and facilities (one major facility taking the load, along with minor facilities) is meant to mirror the real data used under Pilot 2 of the Track & Know Project.

Real postcodes, from Royal Mail, are used to generate the appointment population, real facilities are used based on the British Lung Foundations study of Obstructive Sleep Apnoea, and real GP's are used based on public data from the NHS.

2. How to download the data

The dataset is under a **GNU General Public License v3.0.** Permissions of this strong copyleft license are conditioned on making available complete source code of licensed works and modifications, which include larger works using a licensed work, under the same license. Copyright and license notices must be preserved. Contributors provide an express grant of patent rights.

The data is available publicly via Github.

https://github.com/ibadkureshi/tnk-pilot2data

3. Disclaimer

All data is randomly generated and changes with every run of the generation code. Any resemblance to actual events or locales or persons, living or dead, is entirely coincidental.

This is synthetic artificial data meant for research and tool validation purposes. It cannot and should not be used to make business decisions.

This data differs from the original data in that postcodes from the catchment are randomly selected. There were patterns in the occurrences of OSA and the referral

schedules within the real data that are not reproduced in this synthetic dataset. These patterns are not reproduced because reapplication of this code (w/ the pattern reproduction) on the Track&Know pilot catchment area *could* lead to an exposure of actual patients.

4. Dataset creation process

A Python-3 Jupyter Notebook is bundled with this repository that can be used to recreate the dataset and it breaks down exactly each step.

OpenRouteService is required to 'reconstruct' the imaginary journeys been the patients home and the GP's, and back, and from the patients home to the clinic, and back.

5. Content

Patient Appointment Log:

1. patientno - integer variable giving the 'patient' number. From 1 till maxPatients (10000). Just used to track an appointment of journey.
2. postcode - a UK standard postcode typically of the style AA9 9AA or AA99 9AA. These are real postcodes that actually exist and can contain between 1-100+ domiciles. This is a closed system so further information cannot be included here. The code randomly selects a postcode to represent a 'patient'. A postcode can be selected more than once as there can be more than 1 'patient' per postcode. For the synthetic dataset the catchment area included 48000+ valid postcodes out of which a selection was made 10000 times.
3. latitude - The latitude for the above postcode. This is real and sourced from the public data available on postcodes.
4. longitude - The longitude for the above postcode. This is real and sourced from the public data available on postcodes.
5. gpdate - The synthetic date the in dd/mm/yyyy format for when the 'patient' met their local general practitioner. Synthetically generated based on working days and average 'patients' per day. Bundled data covers a period from 01/01/2019 till 31/12/2019 giving a pool of 260+ possible dates removing weekend. Given a population of 10000 that averages to just less that 40 appointments/day. So, the first 40 get the date 1/1/2019, the next 40 get 2/1/2019, and so on.
6. gpname - Based on public data of registered GP's this practice is selected based on geographical proximity to the 'patients' postcode. The geographic catchment area for the bundled data include 400+ GP practices. These are then assigned to the 10000 'patients' based on distance. It is not compulsory

that every GP is assigned. Given that that <25% of postcodes feature in sample dataset and that a GP covers many adjacent postcodes they are likely many not included.

7. gppostcode - From the public data the real postcode of the registered GP.
8. gplatitude - The real latitude of the postcode of the registered GP (based on a join with the public postcode data). This point is not guaranteed to exactly fall on the practice but will be close by.
9. gplongitude - The real longitude of the postcode of the registered GP (based on a join with the public postcode data). This point is not guaranteed to exactly fall on the practice but will be close by.
10. clinicdate - A synthetic and random date that is based on the gpdate variable and observed behaviour from Track&Know's pilot. clinicdate will fall anywhere between +7 days to +6weeks (42 days) from the GP referral date. The code chooses a random number from 7-42.
11. clinicname - Based on the public BLF open dataset on Obstructive Sleep Apnoea. Manually a list was created to include OSA diagnostic facilities in the geographic area of interest. The code then randomly assigns one of the clinics to each 'patient'. The random selector is biased towards one facility to mimic the real-world behaviour of 1 primary facility supported by secondary/out-reach facilities.
12. clinicpostcode - The real postcode of the facilities as gathered from BLF.
13. cliniclatitude - The real latitude of the postcode of the facility GP (based on a join with the public postcode data). This point is not guaranteed to exactly fall on the facility but will be close by.
14. cliniclongitude - The real longitude of the postcode of the facility (based on a join with the public postcode data). This point is not guaranteed to exactly fall on the facility but will be close by.

Patient Journey Log:

1. patientno - integer variable giving the 'patient' number. From 1 till maxPatients (10000). Just used to track an appointment of journey.
2. epoch – timecode of point on journey
3. latitude – latitude at epoch time
4. longitude – longitude at epoch time
5. startlat – originating latitude for this journey
6. startlon – originating longitude for this journey
7. endlat – ending latitude for this journey
8. endlon – ending longitude for this journey

Demand File:

1. latitude
2. longtiude

3. patientno