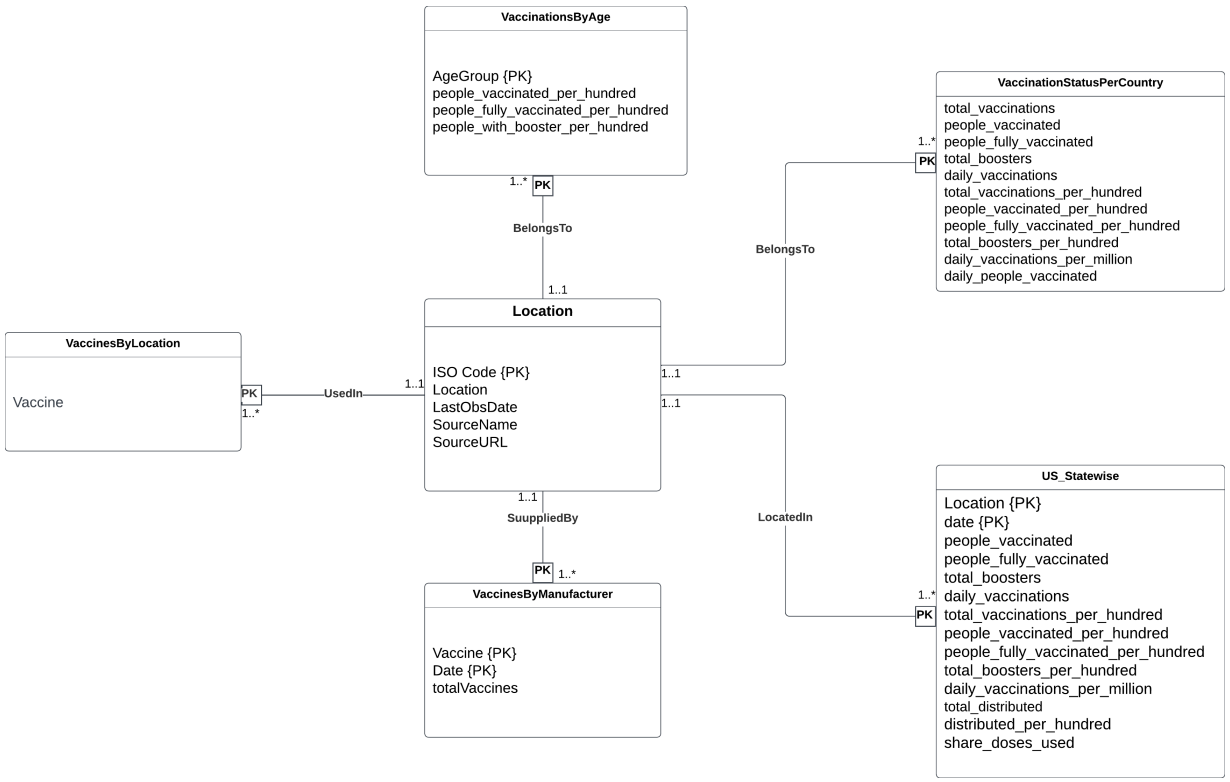
# Assignment 4: Database Design Project

Anandh S - S3976934

## Part B: Designing the database

### Entity-Relationship Diagram




### Assumptions made:

- The 4 CSV files we have for Australia, New Zealand, England and the US are already a part of the main Vaccinations file. It may appear as though these files/data is redundant, however, these tables contain additional data. If we were to create a separate entity, we'll still have to merge this entity with the VaccinationsStatusPerCountry entity as they'd share the same composite primary key. The challenge would be to transfer the additional data in the country/*.csv files into the main vaccinations.csv file (or their respective entities) and to address the missing columns.

- Some attributes can be derived from other attributes within the tables. For instance, the vaccinations csv file has an attribute called total_vaccinations. It can be derived from people_vaccinated, people_fully_vaccinated and total_boosters.

- "daily_vaccinations_raw" was dropped/removed as this is redundant. The source of the dataset also advices against using this column for any purposes.

### Normalisation Challenges:

- The CSV files are not in a format ideal for creating a database. There's plenty of redundancy and repeating groups. For instance, the location names repetitive in all the files. This is not necessary when we have the ISO Code for each location. To handle this, the locations file was split into two tables, Locations and VaccinesByLocation. The locations table would act as a

catalogue that'd have location name and source information for each ISO Code and the rest of the tables will borrow the ISO Codes from this table.

- Vaccines in the Vaccinations file are present as multi-valued attributes for each country. This table was converted to long-format using Excel's Power Query. By removing duplicates in the resulting table, we obtain a table contain the names of the vaccines used in each country.

- As mentioned in point #2 of the previous section, attributes that are functionally dependent on non-primary key attributes are removed from the file as they can be computed using the non-primary key attributes they depend on. However, upon analysing the data, there seems to be a lot of inconsistency as the sum of the three attributes do not always add up to the total_vaccinations values that are present in the file. With an issue like this on our hands, it'd not be ideal to go ahead with a computed attribute for total_vaccinations as we may end up with values that don't match what the sources provide. Inserts and updates on this column would also fail unless the underlying values are updated.

**Database Schema:**

**Location** (LocationName, ISO_Code, LastObsDate, SourceName, SourceURL)

**VaccinesByLocation** (VaccineName, ISO_Code*)

**VaccinesByManufacturer** (VaccineName, ISO_Code*, Date, totalVaccinations)

**VaccinationsByAgeGroup** (ISO_Code*, AgeGroup, people_vaccinated_per_hundred, people_fully_vaccinated_per_100, booster_per_100)

**VaccinationStatusPerCountry** (ISO_Code*, date, people_vaccinated, people_fully_vaccinated, total_boosters, daily_vaccinations, total_vaccinations_per_hundred, people_vaccinated_per_hundred, people_fully_vaccinated_per_hundred, total_boosters_per_hundred, daily_vaccinations_per_million, daily_people_vaccinated)

**US_Statewise** (StateName, date, people_vaccinated, people_fully_vaccinated, total_boosters, daily_vaccinations, total_vaccinations_per_hundred, people_vaccinated_per_hundred, people_fully_vaccinated_per_hundred, total_boosters_per_hundred, daily_vaccinations_per_million, total_distributed, distributed_per_hundred, share_doses_used

## References

‘covid-19-data/public/data/vaccinations at master · owid/covid-19-data’ GitHub, viewed 1 May 2023, <https://github.com/owid/covid-19-data>.

‘How to Create a Computed Column in SQLite’, viewed 9 June 2023, <https://database.guide/how-to-create-a-computed-column-in-sqlite/>.

‘Using Computed Columns in SQL Server with Persisted Values’, viewed 9 June 2023, <https://www.mssqltips.com/sqlservertip/1682/using-computed-columns-in-sql-server-with-persisted-values/>.

Strachnyi, K 2021, ‘The Ultimate Cheat Sheet on Tableau Charts’, Medium, viewed 9 June 2023, <https://towardsdatascience.com/the-ultimate-cheat-sheet-on-tableau-charts-642bca94dde5>.