| Data Integratie  07-06-2022 | Gijsbert Keja,  Femke Nijman  en  Kim Bredenkamp  BIN-3d |
| --- | --- |

# 

# Inhoudsopgave

[**Inhoudsopgave**](#_heading=h.sgxlkpu2pa5q) **1**

[Inleiding](#_heading=h.30j0zll) **2**

[Doel](#_heading=h.dyhon1u0pw62) 2

[Gebruiker](#_heading=h.vf1z0wqmixpd) 2

[Data](#_heading=h.d9oj87f7otdf) 2

[Extract data](#_heading=h.1pacgucfvph7) 2

[Werking van de applicatie](#_heading=h.1fob9te) **3**

[Database opzet](#_heading=h.h5sfk4qxu6mx) 3

[**Software architecture**](#_heading=h.k6hkx35ipwnk) **7**

# Inleiding

## Doel

Het doel van deze applicatie was om patiëntinformatie die momenteel in tekstformaat is in te voegen in een database zodat deze kan worden gekoppeld aan andere databases. Hierdoor wordt een netwerk gecreëerd van patiëntgegevens die gekoppeld zijn aan elkaar. Dit zorgt voor een geordend en beter toegankelijkheid van patient data tussen onderzoeksinstellingen.

## Gebruiker

Deze applicatie is bedoeld om te worden gebruikt door medisch gecertificeerde personen voor de analyse van gegevens om patiënten met een ziekte te diagnosticeren en te behandelen.

## Data

Voor deze applicatie zijn de volgende bestanden gedownload van <https://personalgenomes.ca/data> data repository: PGPC-6.pdf, PGPC-21.pdf en PGPC-52.pdf. Deze bestanden bevatten genomische en klinische data van 3 participanten die hun gegevens meegeven aan het Personal Genomes Project (PGP). Alle pdf bestanden bevatten de volgende informatie: een deelnemersprofiel, aandoeningen en symptomen, inentingen, allergieën, medicijnen en relevante medische procedures. Deze pdf bestanden werden ingelezen met de tabula functie van Python en in een lijst gezet.

## Extract data

De observationele gezondheidsgegevens, de deelnemersprofiel; en aandoeningen en symptomen data, zijn geëxtraheerd uit de pdf-bestanden.

Eerst werden de bijbehorende vcf.gz-bestanden gedownload. In deze bestanden zijn de headers en regels waarin chromosoom 21 voorkomt opgeslagen in een lijst. Tijdens het zoeken, zijn er geen vcf.gz-bestand gevonden voor patiënt PGPC-21.pdf, dus alleen de volgende bestanden zijn gevonden voor deelnemersbestand PGPC-6.pdf en PGPC-52.pdf: PGPC\_0006\_S1.flt.vcf21 en PGPC\_0052\_S1.flt.vcf21.

Om de varianten van het single nucleotide polymorphism (SNP) voor chromosoom 21 te annoteren, werd SNPeff gebruikt. De volgende opdracht is in de terminal uitgevoerd om alleen varianten te annoteren die onder de CDS vallen: “java -jar snpEff.jar GRCh37.75 -no-downstream -no-intergenic -no-intron -no-upstream -no-utr -s - uitgebreid ../chr21.vcf > ../chr21\_ann.vcf”. De resultaten zijn vervolgens gefilterd door missense varianten en frameshift varianten te zoeken in de bestanden en de resultaten op te slaan in de volgende bestanden: 0006.chr21.snpEff.vcf en 0052.chr21.snpEff.vcf. Vervolgens, werden deze bestanden overheen geitereerd met een pythonscript om de top 10 resultaten willekeurig te selecteren en in een lijst op te slaan.

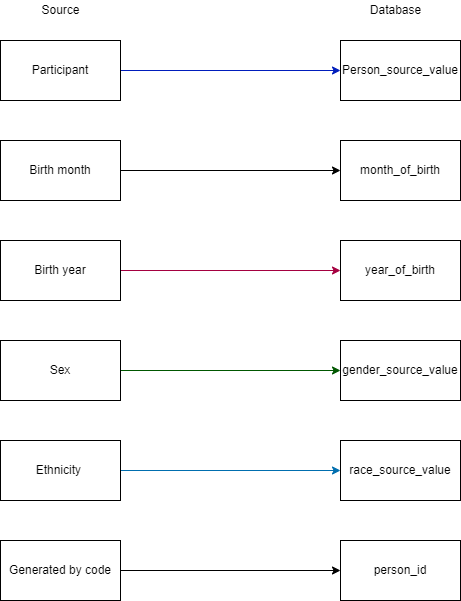
# Werking van de applicatie

Voor de applicatie werd Python versie 3.9 en PostgreSQL versie 14.3 gebruikt. Dit applicatie bestaat uit drie Python scripten: snpeff.py die informatie van de patienten inleest en die informatie over chromosome 21 ophaalt uit de vcf bestanden om te verwerken; postgres\_connector.py maakt een verbinding met de PostgreSQL database en vult de database met de opgehaald informatie over de drie patienten; en workflow.py die de andere twee scripten aanroept en die vervolgens uit te voeren.

## Database opzet

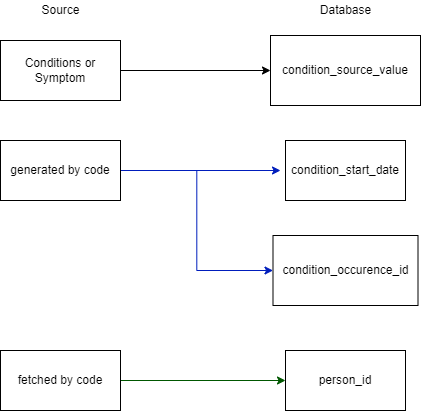
De database die voor dit project is gebruikt is gevuld met Data Grip en PostgreSQL. Er is een koppeling gemaakt met de PostgreSQL database met het Observational Medical Outcomes Partnership (OMOP) Common Data Model (CDM) ingeladen (versie 5.3.1). Er is een verbinding gemaakt met de PostgreSQL database met behulp van de volgende informatie: “postgres.biocentre.nl” servernaam, “DI\_groep\_2” gebruikersnaam, “blaat1234” wachtwoord en “DI\_groep\_2” schema naam.

Vervolgens werden de pdf en vcf.gz/vcf bestanden gebruikt om de database te vullen. Eerst werden de deelnemersgegevens uit de drie pdf-bestanden ingevoegd in de volgende kolommen van de persons tabel: de deelnemer naam in person\_id en person\_source\_value, geboortemaand in month\_of\_birth, geboortejaar in year\_of\_birth, geslacht in gender\_source\_value, etniciteit in race\_source\_value en een computer gegenereerde referentie-ID in person\_id (figuur 1).



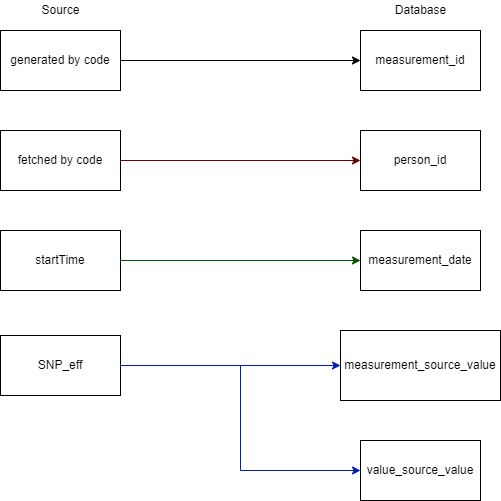
*Figuur 1: Een diagram die het vullen van het persons tabel in MySQL weergeeft.*

Vervolgens, werden de aandoeningen en symptomen gegevens uit de drie pdf-bestanden ingevoegd in de volgende kolommen van de tabel: de deelnemer aandoeningen en symptomen in condition\_source\_value, een computer gegenereerde referentie-ID in condition\_start\_date en condition\_occurence\_id, en een cijfer-ID verzameld uit de bestanden in person\_id (figuur 2). Er is niets toegevoegd aan de condition\_source\_value omdat er geen gegevens voor werden gevonden.



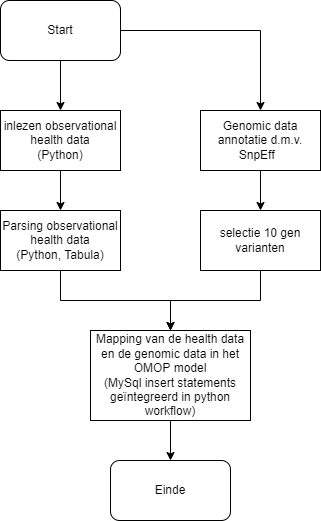
*Figuur 2: Een diagram die het vullen van het conditions tabel in MySQL weergeeft.*

Daarnaast, werden de data uit de SNP\_eff bestanden ingevoegd in de volgende kolommen van het tabel: een random gegenereerde cijfer in measurement\_id, een cijfer-ID verzameld uit de bestanden in person\_id, de begin tijd in measurement\_date, niets in measurement\_concept\_id, niets in value\_as\_concept\_id, SNPeff informatie in value\_source\_value, en SNPeff informatie in measurement\_source\_value (figuur 2). Er is niets toegevoegd aan de measurement\_concept\_id, value\_as\_concept\_id en value\_source\_value omdat er geen gegevens voor werden gevonden in de bestanden.



*Figuur 3: Een diagram die het vullen van het measurement tabel in MySQL weergeeft.*

# Software architecture



*Figuur 4: Een flowchart wat de volgorde van de workflow weergeeft van het programma.*