Supplementary material

Readme

Overview of Finnish national patient data repository for research on medical risk assessment

Table of Contents

[Introduction 3](#_Toc154123379)

[A. CDA R2 Views in Kanta PDR 3](#_Toc154123380)

[B. ICD-10 diagnose group codes in Kanta PDR 3](#_Toc154123381)

[C. ICPC2 codes in Kanta PDR 3](#_Toc154123382)

[D. Physiological measurements unique patients yearly in Kanta PDR 3](#_Toc154123383)

[E. Physiological measurements occurrences per patient in Kanta PDR 4](#_Toc154123384)

[F. Kuntaliitto laboratory studies in Kanta PDR 4](#_Toc154123385)

[G. Risk information in Kanta PDR 4](#_Toc154123386)

[H. Vaccinations in Kanta PDR 4](#_Toc154123387)

[I. Procedures in Kanta PDR 4](#_Toc154123388)

# Introduction

This supplementary material provides data insight of detailed occurrences different data types presented in original paper. Statistics are based on Jasmine project dataset from Kanta patient data repository (PDR) collected 06/2022. Dataset contained pseudonymized CDA R2 documents recorded to the Kanta PDR between 2012 and 06/2022. Individuals were selected for the dataset randomly from whole Finland. Only exclusion criteria for the data collection were that documents were collected only from the time when the person had been over 18 years old. Totally data was collected from 192 399 individuals. Due to plans

for utilizing data for machine learning, the dataset was split randomly to equally sized development and validation datasets.

This study only utilizes the development dataset, ***thus all statistics presented in this supplementary material are based on medical histories of 96 200 patients***.

# A. CDA R2 Views in Kanta PDR

This document contains CDA R2 documents views structures occurrences between 2014 and 2021. The analysis covered all documents which occurred in the N=96200 dataset between 2014-2021.

# B. ICD-10 diagnosis group codes in Kanta PDR

This document contains ICD-10 code group occurrences. The analysis covered all documents which occurred in the N=96200 dataset without limiting to any specific year span. "Total number of diagnoses" is the total amount of found ICD-10 codes in that group including duplicate occurrences among the patient history. "Unique persons who have at least one diagnosis from the group" represents the number of found individuals who have at least one diagnosis during the history which belongs to that group.

# C. ICPC2 codes in Kanta PDR

This document contains ICPC2 codes occurrences. The analysis covered all documents which occurred in the N=96200 dataset without limiting to any specific year span. "Total number of diagnoses" is the total amount of found ICPC2 codes including duplicate occurrences among the patient history. "Number of unique persons who have the diagnosis" represents the number of found individuals who have at least one diagnosis during the history.

# D. Physiological measurements unique patients yearly in Kanta PDR

This document contains occurrences of structurally recorded physiological measurements between 2014 and 2021. The analysis covered all documents which occurred in the N=96200 dataset between 2014-2021. Data describes how many unique patients have those specific measurements recorded into the Kanta at that specific year.

# E. Physiological measurements occurrences per patient in Kanta PDR

This document contains structurally recorded physiological measurements occurrences for individuals. The analysis covered all documents which occurred in the N=96200 dataset. Data describes how many measurements patients have during the history for each specific physiological measurement.

# F. Kuntaliitto laboratory studies in Kanta PDR

This document contains structurally recorded Kuntaliitto laboratory studies occurrences between 2015 and 2021. The analysis covered all documents which occurred in the N=96200 dataset between 2015-2021.

# G. Risk information in Kanta PDR

This document contains structurally recorded clinical risk factor information occurrences between 2014 and 2021. The analysis covered all documents which occurred in the N=96200 dataset between 2014-2021.

# H. Vaccinations in Kanta PDR

This document contains structurally recorded vaccination information (ATC-code, VNR-codes, and protection types) occurrences. The analysis covered all documents which occurred in the N=96200 dataset.

# I. Procedures in Kanta PDR

This document contains procedures occurrences between 2014 and 2021. The analysis covered all documents which occurred in the N=96200 dataset between 2014 and 2021.