FinJeHeW Finnish Journal of eHealth and eWelfare

SCIENTIFIC PAPERS
VERTAISARVIOITU
KOLLEGIALT GRANSKAD
PEER-REVIEWED
www.tsv.fi/tunnus

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

Viljami Männikkö[1,2], Klaus Förger[2], Henna Kujanen[2], Jani Tikkanen[3], Simo Antikainen[2], Joona Munukka[2]

[1] Tampere University (TUNI), Tampere, Finland; [2] Atostek Oy, Tampere, Finland; [3] Oulu University Hospital, Oulu, Finland

**Viljami Männikkö, Atostek Oy, Hermiankatu 3, FI-33720 Tampere, FINLAND. Email: viljami.mannikko@atostek.com**

## Abstract

The Kanta Patient Data Repository (PDR) contains healthcare data from the population of Finland for more than a decade. The repository is a continuously expanding real world dataset produced by many information systems and healthcare service providers. Kanta data has been available for secondary uses such as scientific research since 2019. The data can be requested from the Finnish authority Findata. However, before a request has been accepted, it is difficult to assess if the accumulated data allows answering a specific research question. Publicly available descriptions of data structures in the Kanta PDR do not tell how much they are used in practice. This publication enables future data use cases by providing a view on the overall availability of types of structured health data in the Kanta PDR based on a sample of 96 200 medical histories of over 18-year-old patients. We conclude that the Kanta PDR is a promising source of real world data for development and evaluation of medical risk calculators within the Finnish population. The wide coverage of the Finnish population and timeliness of the data are its strengths as a source of research data also outside of Finnish context. However, the limitations on data availability in variable level need to be considered on a case-by-case basis. Main challenges in the use of data in the Kanta PDR are multiple code systems for laboratory results, short durations of recorded data for specific data types, and missing or very rarely used structured format e.g., in cases of tobacco and alcohol use.

**Keywords:** big data, health and wellness sector, health data, risk assessment, statistics

## Introduction

This paper presents an overview of the structured healthcare data available from the Finnish Kanta Patient Data Repository (PDR). The main motivation is to facilitate development and validation of medical risk prediction. This requires data from many persons over several years. For example, cardiovascular risk calculators have been developed with datasets ranging from 5 209 to 2.4 million people, and with prediction time frames between 5 and 10 years [1]. The dataset must also contain enough cases where the predicted risk has been realized, and suitable input variables. Recording of patient data to the Kanta PDR is mandatory for all public healthcare services providers, and for private healthcare using electronic records including both primary care and inpatient care [2]. Storing data in the Kanta PDR started in 2014, and by the end of

2023, it contained healthcare data from 6.7 million people [3]. As the Kanta PDR contains real world data (RWD), it allows evaluation of the cost-effectiveness of risk calculators, a criterion raised in a previous study [4]. Data from the Kanta PDR can be merged with data from other sources such as genome data from Finnish biobanks. The next sections of the paper give background information on the Kanta PDR, a description of the dataset containing 96 200 medical histories, and an analysis of the availability and history of the main data types offered for research.

***Data in the Kanta PDR***

In this study, we focus on the structured medical records in the Kanta PDR. Kanta Services also include social welfare data, personal health records, and prescription data which all are excluded from this study. The medical records include both structured information and unstructured notes. After late 2014, all public healthcare service providers have had to record patient data to the Kanta PDR [5]. In practice, the Kanta PDR has been taken into use gradually over several years [6].

The data content in the Kanta PDR is similar to health registries of other Nordic countries or other healthcare registries in Finland [7]. In Finland there are many other healthcare registries that are either local registries containing data produced only by, for example, one wellbeing services county, or statistical registries like Care Register for Health Care (Hilmo) [8,9]. Functionally, the Kanta PDR can be viewed as a national health information exchange between service providers, and it also allows access to personal health information through My Kanta Pages [10]. This differs from many other Nordic registries that are mainly administrative allowing, for example, health care surveillance [7]. Previous studies on research potential of Nordic health registries cover the quality of the data on a general level [7,11,12]. In contrast, this study aims to provide data at the level of individual variables.

Health information in the Kanta PDR is split into two groups; key health information and other application area specific health information [13]. The key health information as defined by Finnish authorities contains essential patient information for the provision of healthcare. In practice, it includes diagnoses, clinical risk factors, procedures, physiological measurements, laboratory tests, vaccinations, and imaging studies. Application area-specific data contains, for example, information related to optometry and oral health care. This study focuses on key health information, because it creates a general basis for medical risk assessment.

Health data from the Kanta PDR for research purposes is delivered by default in comma-separated values (CSV) format for a requested set of structured variables [14]. However, all the patient data recorded to the Kanta PDR are stored as XML documents using Health Level Seven (HL7) Clinical Document Architecture Release 2 (CDA R2) format [15]. Also, there are Finland-specific implementation guides and extensions to the HL7 CDA standard [16]. We requested documents in their original format (CDA R2) as our goal is to advance general utilization of health data.

***Kanta PDR for research projects***

Since 2019, data in Kanta has been available for secondary uses such as scientific research, under the Finnish Act on Secondary Use of Health and Social data [17]. We have found few research publications that have used the Kanta PDR as the main data source. For example, data from the Kanta prescription centre has allowed creating a research cohort based on medications [18]. Similar evaluation of the Kanta PDR as we present has been done in the

Valtava project, but the focus was only on diabetes [19].

The Finnish authority Findata issues permits for secondary use of Kanta data and is responsible for facilitating the picking and pseudonymization of the requested data [14]. In practice, the data is extracted by the controller of a registry which for Kanta PDR is Kela, the Social Insurance Institution of Finland. All the requested information must be well justified by the planned research as by default all the information in the Kanta PDR is private. Correctness of data requests is important as the process can be long. In 2024, estimates for processing time of data permits are from 3 to 5 months and for data picks from 3 to 6 months [20].

Data Resource Catalogue contains descriptions of several Finnish data sources available for secondary use, including a description for the key health information in the Kanta PDR [8]. It lists the main variables in every category that can be requested but lacks information of the amount of data that is available in each of the categories, and this paper aims to provide that information.

New data are recorded to the Kanta PDR only by healthcare professionals and only when a patient uses healthcare services. This means that the availability of data varies greatly between persons. However, collecting large datasets with fully defined variables is time-consuming and expensive. This has been an issue in many publications. By the year 2015 close to 3 000 scientific studies indexed in pubmed.org suffered of too small sample size to statistically represent the whole national population [21]. The accumulation of new data to the Kanta PDR for previously studied persons does not require resources from the researchers, and requesting updates for a set of patients is possible. The Kanta PDR reflects the use of healthcare services by the Finnish population without limiting to a subset of service providers.

From the point of view of risk prediction, an alternative to the Kanta PDR as a source of data is the national health risk study FINRISK which mapped the health status of the Finnish population between 1972 and 2012. After 2012, the study continued as a part of FinTerveys and Terve Suomi programmes [22]. The study was made every 5 years and sample sizes for each year were between 4 000 - 10 000 patients. The study included persons aged between 25 and 74 years from specific areas of Finland [23].

Both the Kanta PDR and the FINRISK cohorts include laboratory tests and physiological measurements. The FINRISK also contains structured lifestyle data from questionnaires while in the Kanta PDR the same information would be mainly in written text. In the Kanta PDR, most of the data comes from people with health issues, while FINRISK contains data regardless of the health status of the persons. Compared to the Kanta PDR, the FINRISK study has a longer history in collecting health status data of Finnish people.

## Methods and materials

### *Research questions*

The aim of the study was to find out the current state of the structured data stored in the Kanta PDR. In this paper, we covered the overall characteristics which affect the use of the data for research on medical risk assessment. As supplementary material, we offer information on the availability of variables at a detailed level to allow focus on specific cases of research. Our research questions were:

**RQ1** How long histories can be found for persons from the Kanta PDR?

**RQ2** What kind of sample sizes for diagnoses can be found from the Kanta PDR?

**RQ3** What variables are available in practice in structured format in the Kanta PDR?

***Research data description***

This study is part of the project Jasmine where the whole medical history of 200 000 persons was requested from the Kanta PDR. The data was picked by Kela starting from 2014 up to the first half of 2022 and pseudonymized and delivered by Findata to the secure virtual environment Kapseli. The data included randomly picked persons who had used Finnish healthcare services and were at least 18 years old. The documents were requested only from the time when the person had been over 18 years old. We did not require a minimum number of documents or data types per person as we wished to find out the expected amount of data also for healthy persons.

We received data for 192 399 persons due to the above limitations and persons denying use of their data. Due to plans for utilizing the 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 the results are based on medical histories of 96 200 patients. Because the data was requested and delivered in original CDA R2 XML format, it required additional processing to enable statistical analysis. We selected all relevant information required in the analysis from the documents and stored the information in a separate SQLite database. All the results and statistics shown in this paper are based on that dataset.

The 96 200 medical histories contain 31 million documents. On average, there are 150 documents per person. Over half of the persons in our dataset are identified as female (57.80%). There are also 307 persons in our dataset whose gender is undefined or who have transitioned from one gender to another.

Deaths and causes of deaths are not available in our dataset as currently the information is stored outside of the Kanta PDR. This data could be merged with data from the Kanta PDR with a broader research permit.

Views (Näkymä in Finnish) are used for categorization of medical records. Some views, such as neurology and radiology, cover medical specialties, while others can be used to store different certificates or information of laboratory services, for example [24]. Views dedicated for paediatric specialties and adolescent psychiatry are not comprehensively represented in our dataset due to the age limit in our data request. Thus, they were excluded from the statistics. The complete list of views and their annual occurrences can be found in Supplement A: CDA R2 Views in Kanta PDR. This list allows estimating which application areas contain enough data for research.

Part of the views contained no data for several years after 2014. The most frequently used views have data available already starting from 2014 (Fig. 1). Notably, there has been approximately a tenfold increase in the usage of the laboratory view since the year 2019, which is covered in more detail in results.

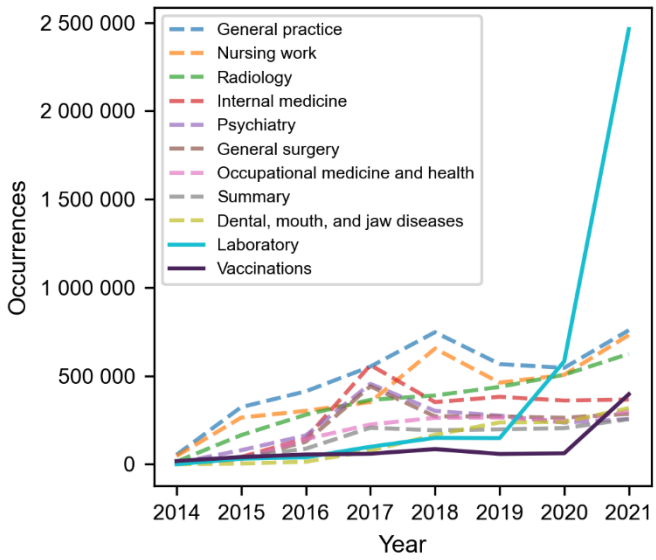


**Figure 1.** Use of ten most common views between 2014-2021.

## Results

### *RQ1: How long histories can be found for persons from the Kanta PDR?*

The majority of persons (79%) have entries from 8 or 9 distinct years, and most people (70%) have 10-50 documents per year. As the Kanta Services are developed in stages, all the data types have different lengths of history, and they have to be considered case by case. The Kanta PDR is not a permanent registry as medical records are only stored 12 years after the death of a person or 120 years from the birth [25].

### *RQ2: What kind of sample sizes for diagnoses can be found from the Kanta PDR?*

The occurrences of diagnoses were collected from the dataset and listed in the Supplement B: ICD-10 diagnosis group codes in Kanta PDR and Supplement C: ICPC-2 codes in Kanta PDR. Both supplements include the total number of diagnoses and the number of persons who had them. To preserve anonymity, exact numbers are provided only for cases with 5 or more occurrences. The ICD-10 (International Classification of Diseases, Tenth Revision) codes were grouped by the first 3 characters as a more detailed level would increase the need for anonymization.

A rough estimate of the total amount of diagnoses for over 18-year-old people in the Kanta PDR can be produced by multiplying the diagnosis count with 47, as the number of people in the sample was 96 200 and in the year 2022, the number of over 18-year-old persons in the Finnish population was 4 537 778 [26]. With this sample size, the order of magnitude of the found diagnoses should be reliable. However, if a new sample would be taken from only one area of Finland, the number of found diagnoses could differ greatly from the overall average. The results for ICD-10 include a total of 1798 different codes from which 1574 were found from at least 5 persons. The results for ICPC-2 (International Classification of Primary Care, 2nd edition) include a total of 1172 different codes from which 922 were found from at least 5 persons.

As diagnoses are often predicted variables in risk evaluation, the expected reliability of the data should also be considered. In principle, the diagnosis data should be reliable as all public healthcare service providers have been obligated to archive data to Kanta after late 2014 [6]. However, from the frequencies of primary diagnoses per year shown in Fig. 2, we can see that the initial growth in the number of recorded diagnoses stopped in the year 2017, and diagnoses are likely to be missing before that year. Thus, we conclude that diagnosis data in the Kanta PDR allows medical risk assessment only after the year 2017. In addition to the primary diagnosis there are also other diagnoses where additional information can be attached. That explains why the frequency of persistence of a diagnosis can be higher than the frequency of primary diagnosis.

An additional exception is diagnosis data from the area of Åland that was found to be missing from the Kanta PDR in the Valtava project [19].

The most frequent ICD-10 code Z01 describes the treatment process and next most frequent ones are common diseases such as respiratory infections (J06) and dental caries (K02). Diagnoses in the Kanta PDR reflect the way healthcare professionals make diagnoses in Finland. Thus, some diseases may be underdiagnosed or diagnosed imprecisely if, for example, their treatment does not require an exact diagnosis. Also, a formal diagnosis may never be added to the Kanta PDR in cases when, for example, a laboratory result alone is sufficient in practice.

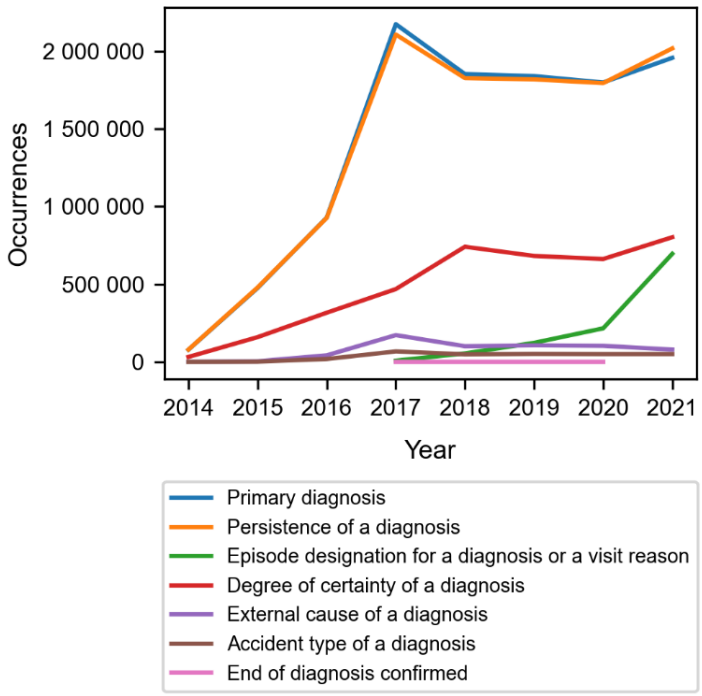


**Figure. 2.** Usage of elements in diagnosis structure and the primary diagnoses between 2014-2021.

***RQ3: What variables are available in practice in structured format in the Kanta PDR?***

*1) Diagnoses*

Diagnoses are recorded in Finland with ICD-10 codes used by doctors and dentists, and with ICPC-2 codes mainly used by other healthcare professionals [24]. The visit reasons reported by the patient are written as free text without the ICD-10 or ICPC-2 codes.

Based on our data, ICD-10 has always been the primary classification system used in the Kanta PDR with a total of approximately 11 million diagnoses while ICPC-2 was used in a total of 1.3 million cases. In addition to the diagnosis code, other structured information can include diagnosis priority, the cause of the adverse effect, diagnosis persistence and diagnosis certainty. From the analysis of the fields in the diagnosis structure (Fig 2), we can see that diagnosis persistence is the most recorded additional information. Diagnosis persistence describes if the diagnosis is permanent or temporary. Diagnosis certainty is recorded in nearly half of the cases and the use of episode identification attachment to diagnosis structure has increased after year 2020. Other structures are found in under 2% of cases.

*2) Physiological measurements*

Physiological measurements include variables such as height, weight, smoking status, blood pressure, and body temperature. 86 types of structural physiological measurements are defined for the Kanta PDR [27], and we found at least one occurrence for 66 types.

Our sample of 96 200 people included approximately 740 000 physiological measurements for around 29 000 individuals. This means that 70% of people did not have any structured physiological measurements after they were 18 years old. There were physiological measurements from 13 distinct years, and 80% of them were from the last three years 2020-2022. The five most common physiological measurements are systolic and diastolic blood pressure, heart rate, body temperature, and weight. Systolic and diastolic blood pressure and heart rate measurements have almost the same number of measurements recorded to the Kanta PDR, because in practice they are received from the same measurement.

Visualization of structured physiological measurements over the years shows that they are rare before the year 2018, and a growth trend after it (Fig 3). The growth matches legislation that requires all healthcare information systems to record all physiological measurements to the Kanta PDR by the end of 2019 [28].

Information on how many and what types of physiological measurements people have is included in the Supplement D: Physiological measurements unique patients yearly in the Kanta PDR. From Fig. 4 we can see that in around 50% of cases, the top 10 most commonly occurring physiological measurements occur only once throughout the patient treatment history. In roughly 40% of cases, measurements occur 2-5 times throughout the history. A complete list of measurement occurrences is included in the Supplement E: Physiological measurements occurrences per patient in Kanta PDR.

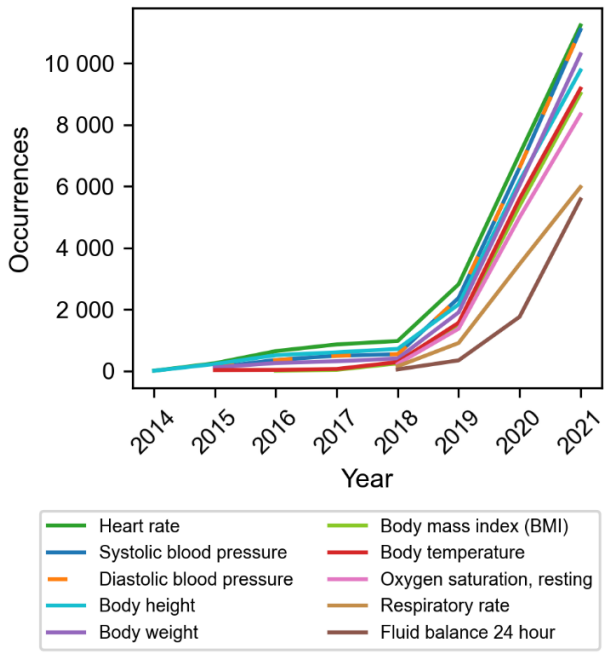


**Figure. 3.** Change of top 10 physiological measurements over time for unique persons with at least one measurement.

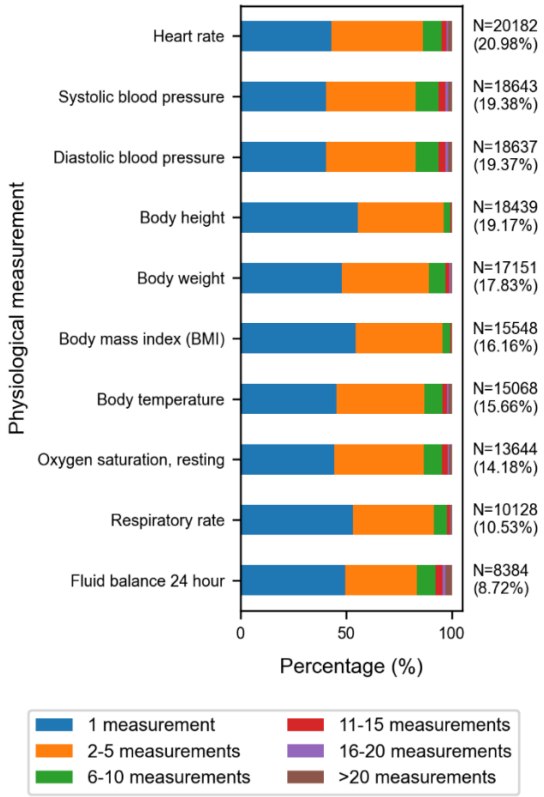


**Figure 4.** Frequency of the top 10 physiological measurements. N and the percentage describe the number of persons who have had at least one physiological measurement of that type during the treatment history in sample set of 96 200 persons.

We conclude that the coverage of physiological measurements in the Kanta PDR is still low but is likely to improve with the current pace of recording of new data. Also, some critical information is missing from the point of view of risk prediction. For example, physiological measurement structure specification is missing for alcohol usage. Also, even though smoking has two structured specifications, they are very rarely used with less than 15 found records.

*3) Laboratory tests*

Laboratory test results are stored in the Kanta PDR under the laboratory view using different code systems and codes. Laboratory service providers can have their own local code systems, which means that results of the same laboratory tests can be stored under different codes and names. In our dataset, we found a total of 318 unique laboratory code systems.

Kuntaliitto laboratory test code system has been developed to standardize the laboratory test codes on a national level. From the total number of recorded laboratory studies in our dataset, the Kuntaliitto code system covered 60% followed by "Fimlab tutkimuskoodisto" at 7%. We also found 35 code systems recorded using Pegasos patient information system with a joint coverage of 21%. This has been reported to be a bug in the Pegasos system [19]. If the bug is fixed, the coverage of Kuntaliitto should rise to 80%, however this will not correct the already recorded data.

The most frequent laboratory results are shown in Fig. 5. They show a growth trend after year 2020, and contain groups such as B-Hbm, E-MCV, B-Eryt, E-MCH, and B-Trom where counts are nearly identical. These laboratory results belong to baseline blood count. A full list of the annual laboratory test occurrences with the Kuntaliitto code system is included in the Supplement F: Kuntaliitto laboratory studies in Kanta PDR.

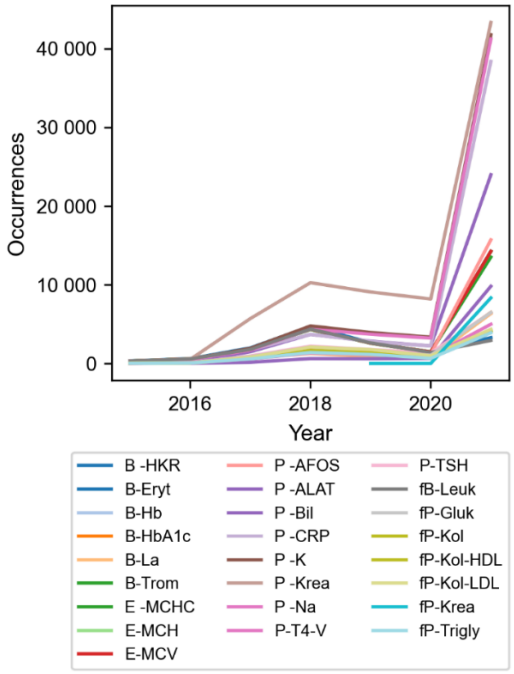


**Figure 5.** The change in the number of most frequently occurring laboratory results between 2014 and 2021.

Overall, the recording of laboratory studies in a structured format is comprehensive. However, the varying code systems and lack of mappings to an international standard limit the usability of the data for research.

*4) Clinical risk factors*

The Kanta PDR contains multiple types of clinical risk factors such as medication reactions & allergies, risk from diseases and treatments, considerations for blood product transfusion, and behavioural risks [24].

The risk information structure contains the level, type, and estimated end date of a risk, related code systems, annotations for patient care, the reason for an end of the risk, and a unique risk identifier [29]. To find out how risk information is recorded in practice, we calculated the occurrences of risk types found in the risk information view. A total of 64 different risk types were found from our dataset. The most common risk type was ”Other infection or need for isolation” that covered nearly 30% of all occurrences. This risk type was used more than any of the other risk types during the year 2021. This outlier was likely related to the rise of COVID-19 cases in that year.

To show development of clinical risk factors over time, we dropped the ”Other infection or need for isolation” out and put the next 10 most common risk types into a timeline (Fig. 6). The graph reveals a growing trend in the recording of clinical risk information. A complete list of yearly occurrences of risk information in our dataset are in the Supplement G: Risk information in Kanta PDR.

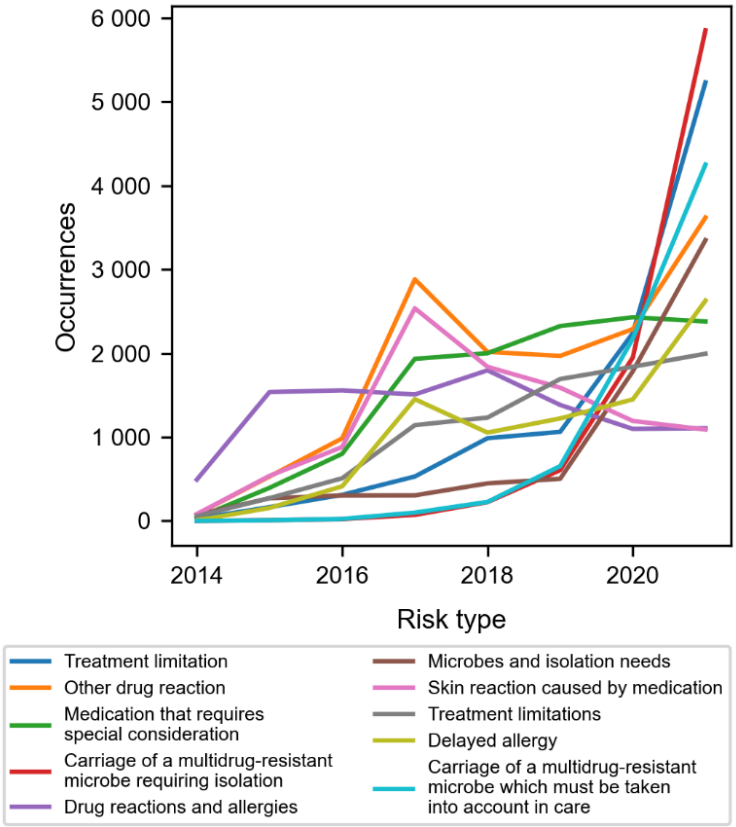


**Figure 6.** Risk type occurrences change over time between 2014 and 2021.

*5) Vaccinations*

Vaccination information has been recorded into the Kanta PDR since 2014 [30], but was often in free text format before 2021 [30]. If vaccination information is available from earlier years, it has been entered retrospectively. According to the Finnish national vaccination programme, most vaccinations are given before the age of 18 [31]. This means that most vaccinations are likely not found from our dataset. However, we can still analyze trends in the use of the structured fields. Structured recording of vaccinations took a big step forward during the COVID-19 pandemic, because it was required from all healthcare service providers.

Over the entire timespan of our dataset, structured information can be found for vaccination ID, vaccine protection, and vaccine information [29]. The vaccine protection identifies the targeted diseases on a general level. The exact vaccination product can be identified with Anatomical Therapeutic Chemical (ATC) codes and Nordic product numbers (VNR codes). Data on adverse effects of vaccines or investigational vaccines were not found in our dataset. Figure 7 shows the number of structured occurrences of ATC codes, vaccine protection, and VNR codes. As we can see, when vaccination information is recorded retrospectively, the protection information is recorded in over 80% of cases, and VNR code or ATC code is recorded only in under 20% of cases. The use of VNR and ATC codes became more common between the years 2010 and 2014, and after 2015 the exact identification of the vaccine products is generally possible. VNR code is almost always available after 2021. A complete list of the occurrences of vaccination protection information, ATC codes, and VNR codes are included in the Supplement H: Vaccinations in Kanta PDR.

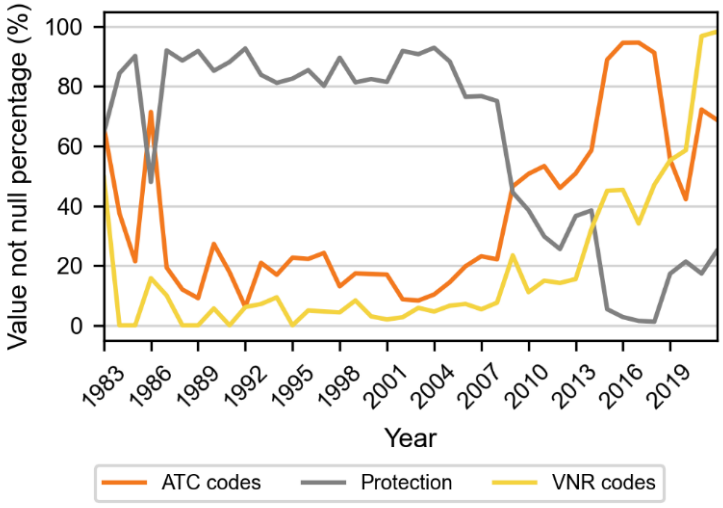


**Figure 7.** Vaccination structure occurrency change between 1983-2022.

*6) Procedures*

The Kanta PDR provides data on healthcare procedures, whether they are performed, planned, or proposed. The data includes details on complications and side effects stemming from completed procedures. However, according to the data resource catalogue operation reports, detailed information on radiology procedures, images obtained from imaging procedures, or any other associated imaging materials are not available for secondary use [8]. In our dataset, there were 1.8 million references to the code system for radiology procedures. As the data resource catalogue states that radiology data are not available for secondary use, we excluded it from our analysis.

To find out which procedures are recorded to the Kanta PDR, we analyzed a subset of the views including only information under specifications THL procedure and THL oral health care procedures. This data includes a procedure code and type and state of the performed procedure. Structures containing complications were not included as they were rare in the dataset. We found that the most common procedures were related to oral healthcare. Over 5 000 unique healthcare procedure types were found. A complete list of procedures and their occurrences are included in Supplement I: Procedures in Kanta PDR.

## Discussion

Our overview on the Kanta PDR is similar to previous reviews of Nordic health registers [7,11,12], but we also provide wider and more detailed lists of the availability of variables as supplementary material to facilitate practical research work. The Kanta PDR is a valuable data source as it holds data practically for the whole Finnish population, and the data can be combined with other sources such as Finnish biobanks. Also, as the Kanta PDR contains real world data, any risk assessment methods that work with the data can be automated in practice.

However, we also found targets for future improvements for the Kanta PDR. Some basic health data, such as structured smoking information and alcohol usage, was found very rarely or not at all. Similarly, many structured physiological measurements such as height and weight were found for less than 20% of persons. Also, the standardization of laboratory results could be better as the standard Kuntaliitto code system was found only in 60% of the entries. These findings are similar to the ones reported previously in the Valtava project [19]. Thus, when considering the use of existing risk calculators [1] in Finland, the suitability of data in the Kanta PDR should be evaluated with the listings of the availability of variables. A concrete improvement for the research use of data in the Kanta PDR would be to extract basic health information such as smoking, height, and weight from text data. Based on our preliminary analysis of textual data, this could radically increase the availability of the variables. Providing results of centralized text mining as variables in the Data Resource Catalogue [18] would save time as researchers would not need to develop methods separately. Also, from a privacy point of view it would be good to develop ways to utilize the data without giving access to raw text data. Another concrete need is a mapping from the multiple laboratory code systems found in the Kanta PDR to an internationally recognized standard. Otherwise, as the Kanta PDR is a continuously growing dataset, simply waiting for more data to accumulate will make it an increasingly useful data source.

## Conclusion

In this study, we found that the Kanta PDR contains many types of structured information that can be used in development of medical risk assessment.

Diagnosis information is widely available for the population of Finland starting from the year 2017. Structured laboratory results and physiological measurements are also promising as we found a large growth in the number of recorded entries after 2020. A few more years of recording will make them valuable inputs for medical risk assessment. We also found that recorded vaccinations have moved from a general specification of given protection to more precise ATC and VNR codes after the year 2015. Other data relevant to medical risk assessment in the Kanta PDR includes clinical risks and medical procedures, which are covered in the supplementary material.

**Acknowledgement**

This work has received funding from Business Finland, and is part of project "Joyful ApparationS of Medical INtelligencE" (Jasmine) which is a subproject of "Agile and Holistic Medical software Development" (AHMED).

## References

[1] Allan GM, Garrison S, McCormack J. Comparison of cardiovascular disease risk calculators. Curr Opin Lipidol. 2014 Aug;25(4):254-65. https://doi.org/10.1097/MOL.0000000000000095

[2] Kanta Services. Legislation [Internet]. Kanta Services; 2023 [updated 2023 Dec 29; cited 2024 Jun 3]. Available from: https://www.kanta.fi/en/legislation

[3] Kela (The Social Insurance Institution of Finland). Amount of data stored in the Kanta Services [Internet]. Kela; 2024 [updated 2024 Mar 1; cited 2024 May 27]. Available from: https://tietotarjotin.fi/en/publication/956744/amount-of-data-stored-in-the-kanta-services

[4] Badawy MA, Naing L, Johar S, Ong S, Hanif AR, Dayangku SN, Chong CL, Tuah NAA. Evaluation of cardiovascular diseases risk calculators for CVDs prevention and management: scoping review. BMC Public Health. 2022;22(1):1742. https://doi.org/10.1186/s12889-022-13944-w

[5] Finlex. Sosiaali- ja terveysministerion asetus terveydenhuollon valtakunnallisista tietojärjestelmäpalveluista [Internet]. Sosiaali- ja terveysministeriö; 2012 Apr 11 [cited 2023 Dec 22]. Available from: https://www.finlex.fi/fi/laki/alkup/2012/20120165

[6] Jormanainen V. Large-scale implementation and adoption of the Finnish national Kanta services in 2010-2017: a prospective, longitudinal, indicator-based study. FinJeHeW. 2018;10(4):381-395. https://doi.org/10.23996/fjhw.74511

[7] Laugesen K, Ludvigsson JF, Schmidt M, Gissler M, Valdimarsdottir UA, Lunde A, Sørensen HT. Nordic Health Registry-Based Research: A Review of Health Care Systems and Key Registries. Clin Epidemiol. 2021 Jul 19;13:533-554. https://doi.org/10.2147/CLEP.S314959

[8] THL (Finnish Institute for Health and Welfare); Statistics Finland; the Data Archives; Sitra. Data Resources Catalogue [Internet]. Data resources catalogue; 2023 [cited 2023 Nov 13]. Available from: https://aineistokatalogi.fi/catalog/studies/3e9d936e-ee2a-4e0e-9344-f b2c85b94e0c

[9] Finnish institute for health and welfare. Care register for Health Care [Internet]. Finnish institute for health and welfare; 2023 [cited 2024 Aug 22]. Available from: https://thl.fi/en/statistics-and-data/data-and-services/register-descriptions/care-register-for-health-care

[10] Ruotanen R, Kangas M, Tuovinen T, Keranen N, Haverinen J, Reponen J. Finnish e-health services intended for citizens - national and regional development. FinJeHeW. 2021;13(3):283-301. https://doi.org/10.23996/fjhw.109778

[11] Bakken IJ, Ariansen AMS, Knudsen GP, Johansen KI, Vollset SE. The Norwegian Patient Registry and the Norwegian Registry for Primary Health Care: Research potential of two nationwide healthcare registries. Scand J Public Health. 2020 Feb;48(1):49-55. https://doi.org/10.1177/1403494819859737

[12] Schmidt M, Schmidt SA, Sandegaard JL, Ehrenstein V, Pedersen L, Sørensen HT. The Danish National Patient Registry: a review of content, data quality, and research potential. Clin Epidemiol. 2015 Nov 17;7:449-90. https://doi.org/10.2147/CLEP.S91125

[13] Kanta-palvelut. Potilastietovarannon määrittelyt [Internet]. Kanta-palvelut; 2024 [updated 2024 May 24; cited 2024 Jun 3]. Available from: https://www.kanta.fi/jarjestelmakehittajat/potilastiedon-arkisto

[14] Findata. Data permits [Internet]. Findata; 2024 [updated 2024 May 3; cited 2024 Jun 3]. Available from: https://findata.fi/en/permits/data-permits/

[15] Dolin RH, Alschuler L, Boyer S, Beebe C, Behlen FM, Biron PV, Shabo Shvo A. HL7 Clinical Document Architecture, Release 2. J Am Med Inform Assoc. 2006 Jan-Feb;13(1):30-9. https://doi.org/10.1197/jamia.M1888

[16] HL7 Finland. HL7 Finland [Internet]. HL7 Finland; 2023 [cited 2023 Dec 12]. Available from: https://www.hl7.fi/

[17] Finlex. Laki sosiaali- ja terveystietojen toissijaisesta käytöstä [Internet]. Sosiaali- ja terveysministeriö; 2019 Apr 26 [cited 2023 Dec 12]. Available from: https://www.finlex.fi/fi/laki/alkup/2019/20190552

[18] Partonen T, Toffol E, Latvala A, Heikinheimo O, Haukka J. Hormonal contraception use and insomnia: A nested case-control study. Sleep Med. 2023 Sep;109:192-196. https://doi.org/10.1016/j.sleep.2023.06.025

[19] Metso S, Tahkola A, Vanhamaki S, Kauppala T, Salo H, Laatikainen T, Salonen J, Veltheim J, Haapakoski J. Valtava-hanke: Diabetesrekisterin pääraportti. Helsinki: Terveyden ja hyvinvoinnin laitos THL; 2022. https://urn.fi/URN:ISBN:978-952-343-846-0

[20] Findata. Finnish Social and Health Data Permit Authority Findata [Internet]. Findata; 2024 [cited 2024 May 27]. Available from: https://findata.fi/en

[21] Andreu-Perez J, Poon CC, Merrifield RD, Wong ST, Yang GZ. Big data for health. IEEE J Biomed Health Inform. 2015 Jul;19(4):1193-208. https://doi.org/10.1109/JBHI.2015.2450362

[22] Terveyden ja hyvinvoinnin laitos THL. Kansallinen FINRISKI-tutkimus [Internet]. THL; 2023 [updated 2023 Dec 19; cited 2023 Dec 12]. Available from: https://thl.fi/fi/tutkimus-ja-kehittaminen/tutkimukset-ja-hankkeet/finriski-tutkimus

[23] Terveyden ja hyvinvoinnin laitos THL. Kansallinen FINRISKI 2012 -terveystutkimus [Internet]. THL; 2019 [updated 2019 Nov 5; cited 2023 Nov 27]. Available from: https://thl.fi/fi/tutkimus-ja-kehittaminen/tutkimukset-ja-hankkeet/finriski-tutkimus/kansallinen-finriski-2012-terveystutkimus

[24] Kauvo T, Virkkunen H. Kirjaamisopas: Potilastiedon kirjaamisen yleisopas: 5.0. Terveyden ja hyvinvoinnin laitos THL; 2022. https://urn.fi/URN:NBN:fi-fe2022031824085

[25] Kanta Services. The data saved in Kanta are shown in MyKanta [Internet]. Kanta Services; 2024

[updated 2024 May 16; cited 2024 Jun 3]. Available from: https://www.kanta.fi/en/data-in-kanta

[26] Statistics Finland. 11rd – Population according to age (1-year 0-112) and sex, 1972-2023 [Internet]. Statistics Finland; 2023 [cited 2023 Nov 27]. Available from: https://pxdata.stat.fi/PxWeb/pxweb/en/StatFin/StatFin__vaerak/statfin_vaerak_pxt_11rd.px/

[27] Terveyden ja hyvinvoinnin laitos THL. FinLOINC - Fysiologiset mittaukset [Internet]. THL Kansallinen koodistopalvelin v7.5.2; 2023 [cited 2023 Nov 24]. Available from: https://koodistopalvelu.kanta.fi/codeserver/pages/classification-view-page.xhtml?classificationKey=273&versionKey=350

[28] Finlex. Sosiaali- ja terveysministerion asetus terveydenhuollon valtakunnallisista tietojarjestelmäpalveluista [Internet]. Sosiaali- ja terveysministeriö; 2015 Oct 7 [cited 2023 Nov 21]. Available from: https://www.finlex.fi/fi/laki/alkup/2012/20151257

[29] HL7 Finland. Kanta – Potilastiedon arkiston Kertomus ja lomakkeet CDA R2-rakenne v 2021. HL7 Finland; 2023. Available from: https://www.hl7.fi/hl7-rajapintakartta/kanta-eark-iston-kertomus-ja-lomakkeet-cda-r2/

[30] Kela (The Social Insurance Institution of Finland). Aineistokatalogi: Rokotukset [Internet]. Kela; 2023 [cited 2023 Sep 22]. Available from: https://www.aineistokatalogi.fi/catalog/studies/3e9d936e-ee2a-4e0e-9344-fb2c85b94e0c/datasets/6905063b-d47f-403d-b92d-ba31551d164c

[31] THL (Finnish Institute for Health and Welfare). Vaccination programme for children and adults [Internet]. THL; 2023 [updated 2023 Dec 7; cited 2024 Jun 3]. Available from: https://thl.fi/en/web/infectious-diseases-and-vaccinations/information-about-vaccinations/vaccination-programme-for-children-and-adults