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

Viljami Männikkö¹, Klaus Förger¹, Henna Urhonen¹, Jani Tikkanen¹, Simo Antikainen¹, and Joona Munukka¹

¹Affiliation not available

February 27, 2024

Abstract
The Kanta Patient Data Repository 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 accessible 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 Kanta 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 patient data repository based on a sample of 96 200 medical histories of over 18-year-old patients. We conclude that 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 Kanta data 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.

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

Viljami Männikkö, Klaus Förger, Henna Urhonen, Jani Tikkanen, Simo Antikainen and Joona Munukka

Abstract—The Kanta Patient Data Repository 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 accessible 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 Kanta 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 patient data repository based on a sample of 96 200 medical histories of over 18-year-old patients. We conclude that 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 of data availability in variable level need to be considered on a case-by-case basis. Main challenges in the use of Kanta data 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.

Index Terms—Data analysis, Data overview, Kanta Services, National Electronic Health Records, Real world data

I. INTRODUCTION

This paper presents an overview of the structured healthcare data available from the Finnish Kanta Patient Data Repository (PDR). This study is part of a project called “Joyful ApparationS of Medical INtellegence” (Jasmine) which is a subproject of “Agile and Holistic Medical software Development” (AHMED) project. The aim of the project Jasmine is to study practical utilization of healthcare data from point of view of medical risk prediction. The main motivation for this overview is to enable planning of research related to development, recalibration, and validation of medical risk prediction. In that context, statistically significant results require a dataset containing a large number of people with several years of recorded medical history. For example, cardiovascular risk calculators have previously been developed with datasets ranging from 5209 to 2.4 million people with prediction timeframes varying between 5 to 10 years [29]. The dataset must also contain enough cases where the predicted risk has been realized, and suitable input variables. Kanta PDR is mandatorily used by all public healthcare services providers, and also in patient information management done in private healthcare including both primary care and inpatient care [32]. As Kanta PDR contains real world data (RWD), it allows evaluation of the cost-effectiveness of risk calculators, a criteria raised in a previous study [30]. It provides identifiable healthcare RWD which can be merged with other data sources containing additional information about individuals. For example, data in Kanta PDR can be merged with data from Finnish biobanks to produce a research dataset with gene information and electronic health records. The next sections of the paper give background information on Kanta PDR, a description of the dataset containing 96 200 medical histories, and then an analysis of availability and history of the main data types that are offered for research.

A. Kanta PDR background

Kanta Services form the Finnish National Health Record System (NHRS) that includes prescription data, medical records, well-being data, and social welfare data [1]. The medical records include unstructured medical notes, given treatments, dental care, laboratory tests, imaging examinations, vaccinations, and medical certificates [2]. Well-being data is health data recorded by individuals themselves by using applications intended for recording to the Kanta Personal Health Records (PHR) [3]. Social welfare data is recorded to the Kanta by social welfare professionals [4]. In this study, we will focus on the medical records recorded in the Kanta PDR. Prescription data is excluded from the dataset due to the limitations of the research permit. PHR data and social welfare data are also excluded because they are new additions and the data coverage is still low [3] [4].

Kanta Services have been developed in stages which means that recording of different data types to the Kanta has started at
different times. Kanta Services entered to production in 2010 with prescription data [6] and in 2012, recording of patient data to the Kanta was started [7]. After late 2014, all public healthcare service providers have had to record patient data to the Kanta [42].

B. Kanta PDR data

Health data from Kanta for research purposes is delivered by default in comma-separated values (CSV) format [5]. CSV format is suitable for research projects where data variables which are needed can be precisely defined in advance. However, all the patient data recorded to the Kanta PDR are stored as Health Level Seven (HL7) Clinical Document Architecture Release 2 (CDA R2) documents. HL7 is an organization for developing standards and frameworks for electronic health records [39]. CDA R2 is a well-defined structured extensible markup language (XML) document format that is used globally in clinical documents [10]. Even though CDA R2 is the global standard for clinical documents, local definitions have to be made for the national healthcare context. In Finland, HL7 Finland and Kela have defined the Finnish CDA R2 document usages and maintain the definitions [40]. CDA R2 documents contain all the clinical information and also meta information like patient, doctor, and system information and links to possible other documents [10]. Before systems can record documents to the Kanta, they have to pass joint testing which is information content-specific. The goal of testing is to test interoperability, security, data protection, and information content-specific requirements. [18] Currently structurally recorded data is the main format offered for the secondary use, and thus the focus of this study. In this study, we analyzed documents in their original format (CDA R2) as our goals include developing general utilization of health data in Kanta PDR instead of focusing on a limited number of variables.

Health information in Kanta PDR is split into two groups; key health information and other application area-specific health information. In this study, we focus on key health information, because it creates general basis for medical risk assessment, and matches the data types offered for research in Data Resource Catalogue [28]. The key health information as defined by Finnish authorities contains essential patient information for the implementation of healthcare. In practice, it includes diagnoses, clinical risk factor information, procedures, physiological measurements, laboratory tests, vaccinations, and imaging studies. Application area-specific data contains, for example, information related to optometry, oral health care, medication labels, first aid, referral and treatment feedback, and appointments. [19]

C. Kanta PDR usage for research projects

Kanta PDR as a data source differs from datasets gathered primarily for research purposes. Kanta PDR has clear advantages, but the secondary nature of research use creates also challenges. Issues separating primary and secondary research data in context of research on risk assessment are process of requesting the data, number of available individuals, length of patient history, data continuity and punctuality, variable coverage over individuals and update cycle of the data. These aspects will be covered next.

From 2010 to 2019, the data in Kanta services was available only for its primary use cases in healthcare. Since 2019, data in Kanta has been accessible also for secondary uses such as scientific research, based on the Finnish Act on Secondary Use of Health and Social data [41]. The data was not immediately accessible as the processes for data requests were under the development. Thus, data in Kanta has been accessible for research for a relatively short period of time when this study has been done (2023). Due to these facts we were able to find quite few research publications which have used Kanta as the main data source. Existing examples include use of Kanta prescription center for creating a research cohort based on different medications [20]. Other examples include reviews of usability of the Kanta [21]. Similar evaluation of Kanta PDR as we present has been done in the Valtava-project [31]. The main difference is that the Valtava-project focused only on data related to diabetes while this paper considers medical risk calculation more widely.

Data for scientific research is requested from the Findata, who is responsible for facilitating the picking and pseudonymization of the requested data [5]. All the requested information must be well justified by needs of the planned research as by default all the information in Kanta PDR is private. The entire data request process can be long especially for large requests. In the case of project Jasmine, the process took 2 years and one month from submitting the data request to having all data accessible.

Data Resource Catalogue has been developed by the Finnish Institute for Health and Welfare (THL), Statistics Finland, the Data Archives, and Sitra [28]. It contains descriptions of several data sources in Finland that are available for secondary use, including a description for the key health information in Kanta PDR. The description lists the main variables in every category that can be requested. The catalogue does not have information of the amount of data that is available in each of the categories, and this paper aims to provide that information.

New data is recorded to Kanta PDR only by healthcare professionals and only when a patient uses healthcare services. This leads to all persons not having the same kind of data from the same time periods. This means that missing variables are the norm which can complicate use of the data. However, collecting datasets with fully defined variables is time-consuming and expensive. This has been an issue in many publications, by the year 2015 close to 3000 scientific studies indexed in pubmed.org suffered of too small sample size to statistically represent the whole national population [22]. The accumulation of new data to Kanta PDR for previously studied persons does not require resources from the researchers, and requesting updates for a set of patients is possible. Kanta PDR reflects use of healthcare services by the Finnish population without limiting to a subset of service providers. It provides RWD from an individual level still maintaining high coverage of the population. One of the strengths of Kanta PDR is the possibility to track changes in the health status of individuals because each document recorded to the Kanta contains patient
From risk prediction point of view, 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 other studies FinTerveys and Terve Suomi [34]. The study was made every 5 years and sample sizes for each year were between 4000 - 10000 patients. It contained persons aged between 25 and 74 and had persons only from specific areas of Finland. [33]

Both Kanta PDR and FINRISK cohorts include laboratory tests and physiological measurements. In Kanta PDR, most of that data comes from people with health issues, while FINRISK contains data regardless of the health status of the persons. Both datasets contain information about lifestyle, however in FINRISK the data is formatted as questionnaire results while in Kanta PDR the same information would be mainly in written text. Compared to Kanta PDR, the FINRISK study has a longer history in collecting health status data of Finnish people.

II. METHODS AND MATERIALS

A. Research questions

The aim of the study is to find out the current state of the data stored in the Kanta PDR and how the data suits to research on medical risk assessment. This study evaluates the authentic CDA R2 patient documents recorded to the Kanta between 2014-2022. Based on identified structures and detailed evaluation, we create an overall picture of data quality for medical risk assessment. We consider three different aspects of the data; sample sizes we can expect, data history length, and available inputs.

In this paper, we aim to cover the overall characteristics which affect the use of the data for research. As supplement material, we offer information on availability of variables on a detailed level to allow focus on specific cases of research. Our research questions are:

RQ1 How long histories can be found for persons from Kanta PDR?
RQ2 What kind of sample sizes for diagnoses can be found from Kanta data?
RQ3 What inputs are available in practice in structured format in Kanta PDR?

B. Research data description

This study is part of Jasmine project where the whole medical history of 200 000 persons was requested from Kanta PDR. Data was picked starting from 2014 and the most recent documents are from the first half of 2022. The included persons were randomly picked from over 18-year-old persons in the Finnish population. Also, documents were requested only from the time when the person had been over 18 years old. We received data for 192 399 persons. 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 the statistics presented in this article are based on medical histories of 96 200 patients.

The 96 200 medical histories contain 31 million documents. On average, there are 150 documents for each person. Over half of the persons in our dataset are identified as female (57.80%). Compared to the gender ratio of Finland, women are overrepresented in our dataset. In 2022, the percentage of female population in Finland was 50.50% [13]. There are also 307 persons in our dataset whose gender is undefined or who have transitioned from one gender to another.

Information related to deaths and causes of deaths are not available in our dataset as currently the information is not stored into the Kanta PDR. Instead, information about deaths are reported in Finland to the Digital and Population Data Services Agency (Digi- ja väestötietovirasto), and cause of death information is archived by Statistics Finland [14]. This data could be merged with Kanta PDR data with a broader research permit.

Fig. 1 shows the birth year distribution of the data set. Based on this, we can see that most medical histories are limited by the start year of the data pick (2014) instead of the person becoming 18 years old.

Views (Näkymät in Finnish) are used for categorization of medical record entries under different entities. Some entities, such as neurology and radiology, are specific to medical specialties, while others can be used to store different certificates or information of laboratory services, for example [11]. The complete list of available views is stored in classification “AR/YDIN - Näkymät” [15]. Views that are dedicated for pediatric specialties and adolescent psychiatry are not comprehensively represented in our dataset due to the age limit that was used in data picking. For this reason, views covering pediatrics, pediatric neurology, pediatric psychiatry, and adolescent psychiatry were excluded from the statistics.

Fig. 2 presents a summary of the usage of different views in our dataset. 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 contain no data for several years after
For example, support for the optometry view was added to Kanta in 2018 [8] and in our dataset, the first entries in it are from 2021. The 10 most frequently used views have data available already starting from 2014 (Fig. 3). The use of all views has increased over the years, as has the overall amount of data that is entered into the Kanta system [9]. 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 section III.

III. RESULTS

A. RQ1: How long histories can be found for persons from Kanta PDR?

The majority of persons have entries from a period of 8-9 years and there are also persons with histories longer than 10 years (Fig. 4). On average, a person has 10-50 documents per year (Fig. 5). 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.

Kanta has data from a relatively short period of time if it is compared to, for example, the FINRISK study where the first study was made in 1972. Currently the life expectancy in Finland is between 70 and 80 years depending on the year of birth [25]. This means that Kanta data covers in general approximately 10% of the expected lifespan of the people. After 10 years the average length of data is approximately 18 years which is then almost 20% of the expected lifespan of the people living in Finland. The dataset will not grow forever due to rules on data storage. Medical records are stored 12 years after the death of a person or 120 years from the birth. For prescription data that is not part of our dataset, the limit on storage is only 22.5 years. [26]

B. RQ2: What kind of sample sizes for diagnoses can be found from Kanta data?

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
Diagnoses in Kanta PDR depend on and reflect the way healthcare professionals make diagnoses in Finland. Table I shows that the most occurring ICD-10 code Z01 describes the treatment process and other 4 the most occurred codes are general diseases in Finland. Certainty and priority of a diagnosis can be represented in their own structures in the Kanta documents as will be introduced later in section III. When using diagnosis information for validation of risk prediction, it should be noted that 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 Kanta PDR in cases when for example a laboratory result alone is sufficient in practice.

<table>
<thead>
<tr>
<th>ICD-10 diagnosis categories</th>
<th>Category description</th>
<th>Number of individuals who have the diagnosis</th>
<th>Overall amount of diagnosis marked to the Kanta</th>
</tr>
</thead>
<tbody>
<tr>
<td>Z01*</td>
<td>Other special examinations and investigations of persons without complaint or reported diagnosis</td>
<td>46950</td>
<td>125346</td>
</tr>
<tr>
<td>J06*</td>
<td>Acute upper respiratory infections of multiple and unspecified sites</td>
<td>40618</td>
<td>136306</td>
</tr>
<tr>
<td>K02*</td>
<td>Dental caries</td>
<td>39661</td>
<td>139249</td>
</tr>
<tr>
<td>M54*</td>
<td>Dorsalgia</td>
<td>37081</td>
<td>156746</td>
</tr>
<tr>
<td>I10*</td>
<td>Essential (primary) hypertension</td>
<td>33108</td>
<td>243298</td>
</tr>
</tbody>
</table>

C. RQ3: What inputs are available in practice in structured format in Kanta PDR?

1) Diagnoses: The use of diagnosis information in Finland is guided by general instructions on recording of medical information. The general instructions define that all diagnoses which occur during the visit or which significantly influenced the treatment should be recorded with ICD-10 or ICPC-2 codes. Furthermore, doctors and dentists are only ones to record diagnoses, while other healthcare professionals record visit reasons to the Kanta. The visit reasons reported by the patient should always be written as free text without using the ICD-10 or ICPC-2 codes. In special care, diagnoses are always recorded as ICD-10 codes, and in primary healthcare, they can be recorded as ICD-10 or ICPC-2 codes. [11]

Based on our data, ICD-10 has always been the primary classification system that is used in Kanta PDR (Fig. 6). In addition to the diagnosis code itself, other structured information can also be recorded in diagnosis structure. This information can include diagnosis priority, the cause of the adverse effect, diagnosis persistence and diagnosis certainty. By analyzing technical CDA R2 structure occurrences within the diagnosis structures (Fig 7), we can see that diagnosis persistence is recorded for almost all primary diagnoses. 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 describe the physiological state of a person with variables such as height, weight, smoking status, blood pressure, and body temperature. These can be used as inputs in medical risk prediction.

Structurally recorded physiological measurements in the Kanta PDR are found under the classification “FinLOINC - Fysiologiset mittaukset” that contains specifications for 86 physiological measurements [36]. In the Kanta PDR sample of 96 200 people, there were around 740 000 structurally recorded physiological measurements for around 29 000 individuals. This means that around 30% of people in the sample have at least one physiological measurement that is...
structurally recorded. In our sample set we have records of physiological measurements from 13 distinct years, where 80% of the measurements are from years 2020-2022.

By analyzing the distribution of all physiological measurements over the different types of physiological measurements (Fig 8) we can see that 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, because all these values are received at the same time from the same measurement.

We also analyzed how the number of physiological measurements has evolved in the sample population over the years. From the results we can see that structurally recorded physiological measurements are rare before the year 2018, and the coverage of individuals starts to grow after it (Fig 9). This suggests that the measurement results were recorded mostly as free text into the patient ongoing treatment report before the year 2018. Act 1257/2015 about requirement of healthcare service providers to record data to Kanta, §3 says that all healthcare information systems must record all physiological measurements to the Kanta at the latest on 31.12.2019 [37]. Based on the statistics, we can conclude that the act has created a significant increase in recording of physiological measurements to Kanta PDR in structured format.

Information on how many and what types of physiological measurements people have is included in the supplement D: Physiological measurements unique patients yearly in Kanta PDR. From Fig. 10 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. For some types of measurements the average number of measurements may be highly influenced by only a small number of people with very high number of measurements during an inpatient episode. A complete list of measurement occurrences is included in the supplement E: Physiological measurements occurrences per patient in Kanta PDR.

Even though structured recording of physiological measurements has been increasing throughout the years, quite a small number of people are still covered. Also, some critical information is missing from risk prediction point of view. For example, physiological measurement structure definition
is missing for alcohol usage that is a key input parameter in many risk prediction models. Also, even though smoking has a structure definition, it is very rarely used.

3) Laboratory tests: Laboratory test results are key information for assessment of medical risk and the health status of individuals. Many diseases and health changes can be identified by laboratory tests. When laboratory test results are stored into Kanta PDR, all healthcare service providers can utilize them when a person is under treatment and it is possible to compare previous laboratory test results with new results.

Laboratory test results are stored in Kanta PDR under the laboratory view using different laboratory test code systems and codes. Laboratory service providers can have their own local code systems, which means that same laboratory test results can be stored under different codes and names. In most cases, laboratory codes follow a naming structure that contains three sections: prefix, study name abbreviation, and back suffix. Typically, the prefix defines the sample type. In the case of blood tests, for example, it defines the blood component. The actual study name abbreviation section can contain, for example, Finnish study long name, Finnish abbreviations or national abbreviations. The back suffix is used to clarify the research entity. [27] For example, in the study name “F-Gluk-O” ‘F’ describes that fasting is required before the study, ‘Gluk’ signifies a glucose (fin. glukoosi) study and ‘O’ means that results are given in verbal format. However, laboratories are not required to use these naming standards or the same national codes for studies. Instead, they can use their own naming styles for each study, which complicates the secondary use of Kanta laboratory data. For example, the search for input variables of a medical risk calculator may require manual interpretation of different test codes and high knowledge of different laboratory tests to understand correctly their differences.

Kuntaliitto laboratory test code system has been developed in an attempt to standardize the laboratory test codes on a national level, but currently all laboratories do not use it. In our dataset, we found a total of 318 unique laboratory code systems. From the total number of recorded laboratory studies in our dataset, Kuntaliitto code system covered 60% followed by “Fimlab tutkimuskooodisto” at 7% as seen in Table II. We also found 35 code systems all named “Pegasos” to have a joint coverage of 21%. Pegasos is a patient information system that is used by multiple different healthcare service providers. Same finding has been done in Valtava-project which concluded that Pegasos used local codes systems instead of the Kuntaliitto in an error, and this was planned to be fixed during year 2022 [31]. After the Pegasos code system issue is fixed, the coverage of Kuntaliitto should rise to 80%, however this will not correct the already recorded laboratory studies. Thus, having definitions of laboratory code systems publicly available for researchers would help interpreting the data found in Kanta PDR.

P-Krea (Creatinine) study has been the most frequently occurring laboratory result throughout the whole history (Fig. 11). It has been significantly more common than other studies and one explanation can be that creatinine level of same persons is checked multiple times when the causes of the symptoms have been clarified. We can also see from the statistics small groups like B-Hbm, E-MCV, B-Eryt, E-MCH, and B-TRom where study counts are around the same. These studies belong to baseline blood count and their high occurrence corresponds to the expected occurrences compared to other studies. A full list of Kuntaliitto code system laboratory study occurrences in yearly level is included in supplement.

![Fig. 10. 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 96200 persons.](image)
Fig. 11. The change in the number of most frequently occurring laboratory results between 2014 and 2021

Overall, laboratory studies are quite well recorded in structured format. The mentioned challenges affect more the data usability for analysis. Currently the data processing requires high level of understanding of the data format and the clinical background of different studies. Data processing and availability could be eased by unifying the used code systems to one national code system with clear descriptions.

4) Clinical risk factor information: Patient information contains a multitude of different types of risk information such as Medication Reactions & Allergies, food allergies, risk diseases and treatments, Considerations for Blood Product Transfusion, Special Considerations in Treatments, Transferred Tissues and Artificial Materials, Treatment Limitations, Treatment Guidelines, and Treatment Deviations and behavioral risks [11]. Risk information can be used widely in medical risk assessment. For example, they can be used as input variables in certain risk models or for creating a specific study cohort.

Risk information structure contain information on estimated end date of risk, code systems related to risk, annotations for considering risk in patient care, reason for risk end, risk level, risk type and unique risk identifier. [16] To find out how risk information is recorded in practice, we calculated the occurrences of risk types found in risk information view. Top 10 most frequently occurring risk types cover nearly 70% out of all 64 different risk types found from our dataset. The most common risk type was "Other infection or need for isolation" that covered nearly 30% of all occurrences in our dataset. Based on a data that risk type occurrence was significantly higher during year 2021 compared to other years and risk informations. COVID-19 pandemic in general caused a high need for isolation. In many cases COVID-19 infections and isolations were recorded under this risk category in Kanta which explains high increase of that risk category in 2021.

To show development of clinical risk factor information development through years we dropped the "other infection or need for isolation" out and took next 10 the most common informations and put them into timeline (Fig. 12). From that development we can see that recording of risk informations has increased evenly through the years. A complete list of yearly occurrences of risk information in our dataset are included in supplement G: Risk information in Kanta PDR.

5) Vaccinations: Vaccination information has been added to Kanta PDR since 2014 [12]. Vaccination information can be available from earlier years, if it has been entered retrospectively. According to the Finnish national vaccination programme, the majority of vaccinations are given before the age of 18 (THL, 2022b). This means that most vaccinations are likely not found from our dataset. However, with our dataset we can still analyze the development of vaccination information recording in structured format because structured format usage highly depends on development of systems recording data to Kanta. Many healthcare information systems recorded vaccinations to Kanta in free text format before 2021 [12]. Vaccination information is available in structured format starting from summer 2021 [12] and it is found either from continuous patient report view under header “Ennaltaehkäisy...
(preventio)” or from a separate view for vaccinations [11]. Recording of vaccinations in structured format took a big step forward during the COVID-19 pandemic, because it was required from all healthcare providers. After that, the structured format became the standard way of recording vaccination data.

In our dataset, the first vaccination entry in structured format is dated 1908. Over the entire timespan of our data, structured information can be found for vaccination ID (code 19.7), vaccine protection (code 19.1), and vaccine information (code 19) [16]. The voluntary field for specification of vaccine protection (code 19.8) that can be filled with free text has also been used. Structures that contain information about adverse effects of vaccines or investigational vaccines are not found in our dataset, which indicates that they are rarely used. Fig. 13 shows the number of structured occurrences of ATC-codes, vaccine protection, and Nordic product numbers (VNR-codes). As we can see, when vaccination information is recorded retrospectively, the protection information is recorded in over 80% of cases. In those cases, VNR-code or ATC-code is recorded only in under 20% of cases. It is possible that those codes have not been available for healthcare professionals at the time of recording of the vaccination information. The biggest change has happened between 2010 and 2014 when protection information occurrence has decreased significantly and the occurrence of VNR-codes and ATC-codes has started to increase. Currently, it seems that almost all new vaccination recordings in Kanta have a VNR-code. It can be used to identify the exact vaccine product and vaccine protection. Based on structure occurrences, some kind of information about vaccine protection is available from a long period of time. Some cases when there is only the protection structure used the exact protection cannot be identified and only free text description is available. Due to the protection description being a free text field, it can contain a variety of descriptions and detailed protection information can be missing. It is safe to say that vaccinations that are structurally recorded after 2015 have either VNR-code or ATC-code recorded, which means that vaccine products and protection information can be identified in more detail. A complete list of the occurrences of vaccination protection information, ATC-codes, and VNR-codes are included in Supplement I: Procedures in Kanta PDR.

6) Procedures: Kanta provides data on healthcare procedures, whether they are performed, planned, or proposed. This includes details on complications and side effects stemming from completed procedures. However, Kanta does not include operation reports, detailed information on radiology procedures, images obtained from imaging procedures, or any other associated imaging materials. [16] In our dataset there were 1.8 million references to radiology procedures code system. Due the Kanta lacks of detailed information about those procedures and Data Resource Catalogue states that radiology procedures are not available for secondary use currently [28], they are excluded from this analysis. Procedure information can be used in study cohort definition and for identification of individuals who have gone through specific healthcare procedures in their treatment history. In risk analysis, information can be used for example in those cases where different procedures affect the predicted risk.

Due procedure code can be used in multiple structures in Kanta we chose to classify a subset of structured information under THL procedure and THL oral health care procedure classifications. This data includes the procedure code and the type and state when the procedure was performed. Structures containing complications were not classified as they did not exist very often in the dataset. The most common procedures relates to oral healthcare which is expected result because many of the Finnish people regularly use oral healthcare services. In total over 5000 different unique healthcare procedures were found in our analysis. A complete list of procedures and their occurrences are included in Supplement I: Procedures in Kanta PDR.

IV. Conclusion

In this study, we found that Kanta PDR contains many types of information that can be used in development of medical risk assessment. Diagnosis information is widely available for the population of Finland starting from year 2017. Structurally recorded laboratory results and physiological measurements are also promising as we found a large growth in amount of recorded entries after 2020. A few more years of recorded medical history will make the found laboratory results and physiological measurements valuable inputs when developing new risk assessment methods. We also found that recorded vaccinations have moved from a general definition of given protection to more precise ATC and VNR codes after year 2015. Other data relevant to medical risk assessment in Kanta PDR includes clinical risks and medical procedures which are listed in more detail in supplementary material of this paper.

We found that the Kanta PDR provides good sample sizes even for many rare medical conditions as the Kanta PDR holds data practically for the whole Finnish population. The value of the data in Kanta PDR is increased by the possibility of combining the data with other sources containing for example social welfare data. Also, as Kanta PDR contains real world data, any risk assessment methods that work with the available data can be automated and taken into practical use in Finland.
However, we also found targets for future improvements in the Kanta PDR. Some basic health data, such as smoking information and alcohol usage, were noticed to be found very rarely or not at all in a structured format. Similarly, some physiological measurements that are relevant for all persons such as height and weight were found in structured format, but are far from covering the whole population. Also, the standardization of code systems used in laboratory results could be better as the standard Kuntaliitto code system was found only in 60% of the entries. Thus, the suitability of data in Kanta PDR to a specific case of risk assessments should be evaluated with the detailed lists on availability of variables found in supplementary material.

A concrete development for improving the research use of data in Kanta PDR would be to extract basic health information such as smoking, height, and weight from text data. Doing the text mining centrally would save time for researchers as there would be no need for everyone to develop the methods separately. Also, from a privacy point of view it would be good to develop ways to utilize the data without having to give raw text data to all researchers. Another concrete need is a mapping between the laboratory code systems found in Kanta PDR as that would unlock the use of laboratory data recorded in local code systems. Otherwise, as Kanta PDR is Kanta PDR as that would unlock the use of laboratory data that would have the most impact in practice. This work has received funding from Business Finland.

ACKNOWLEDGMENT
This work has received funding from Business Finland.

REFERENCES

[28] THL. Data Resource Catalogue. https://aineistokatalogi.fi/catalog/studies/3e9d936e-ee2a-4e0e-9344-fb2c85b94e0c (Accessed 13.11.2023)


Viljami Männikkö Faculty of Medicine and Health Technology, Tampere University, Tampere, Finland

Viljami Männikkö received the B.Sc. and M.Sc. degrees in technical-science from Tampere University, Finland, in 2018 and 2019 respectively. He joined the Faculty of Medicine and Health Technology, Tampere University (TUNI), in 2022, as an Ph.D. student. He joined as a healthcare software designer and project manager at Atostek, in 2018. His research interests include national electronic health record systems, automated health risk prediction and e-health.

Klaus Förger Atostek Oy, Hervanta, Tampere, Finland

Klaus Förger received the Doctor of Technology (Degree) in 2015 from the Department of Computer Science, Aalto University, Finland. He currently works with research related to utilization of medical information at Atostek, Finland.