Secondary use of health data in Kanta for research

Klaus Förger, *D. Sc. (Tech.)*
klaus.forger@atostek.com

21.11.2022

Topics

- Background
- Project Jasmine: AI/ML research related to Kanta
- Why AI is needed with National Health Records?
- Rules and regulations
- Accessing the data in practice
- Outcome so far
- Open research questions

ATOSTEK

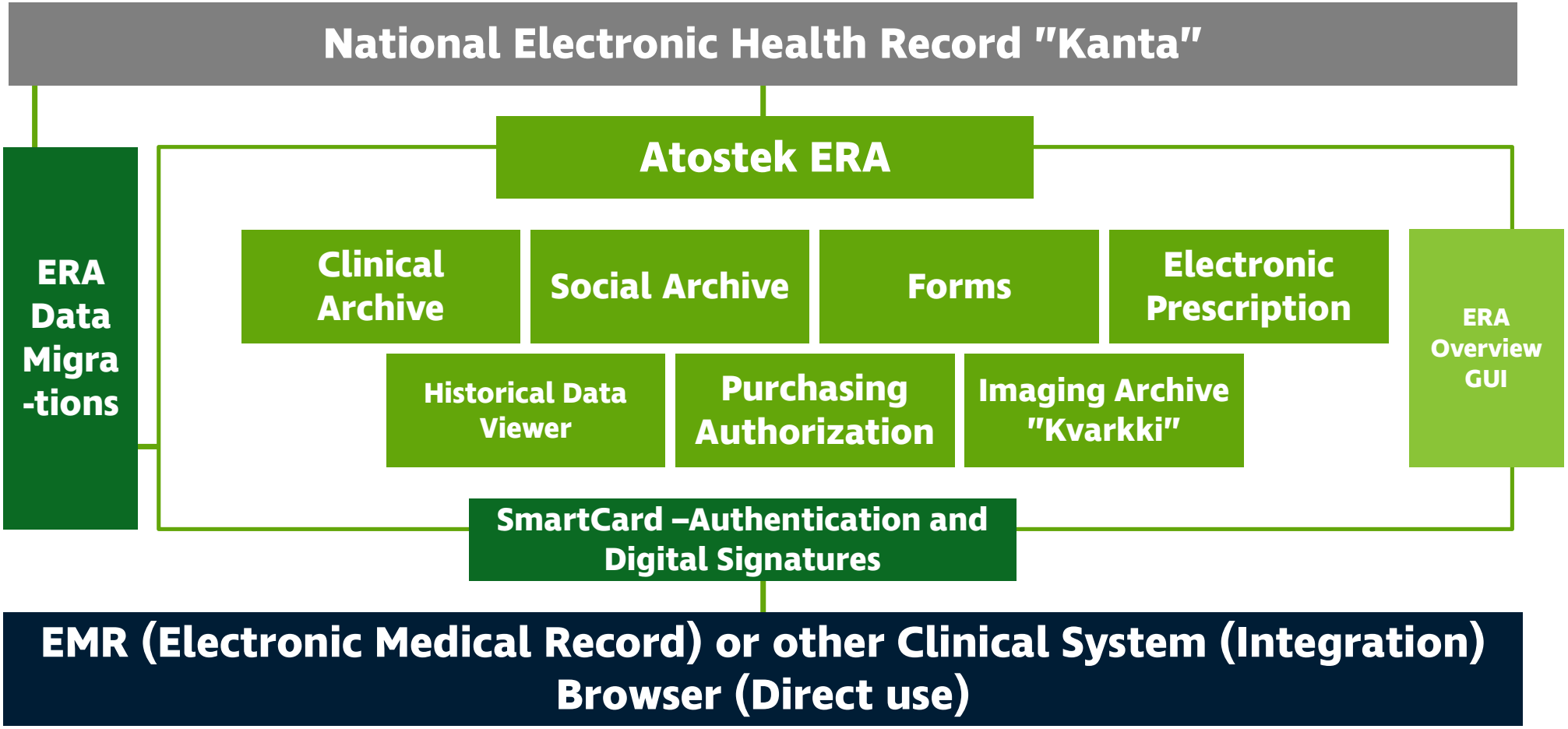
Sectors Atostek works on

Industrial product development

Healthcare and medical applications

Public-sector IT consulting

ATOSTEK



ATOSTEK

ERA services can be used **browser-based** or **integrated** directly into your system. Certification and maintenance have already been taken care of for you.

Project Jasmine

- Jasmine is a research project within the scope of AHMED
  - How can automated analysis of Finnish healthcare data be done?
  - Publish research papers as the data in Kanta for individuals is not available for product development
- We have applied for data from Kanta in original XML format
  - Application process started on 26.3.2021
  - Application accepted on 19.5.2022
  - Data transferred from Kela to Findata on 8.11.2022
  - Data pseudonymization by Findata still in progress on 21.11.2022

ATOSTEK

Why AI is needed with National Health Records?

- The data is not in a user friendly format
- Ideally a Digital Twin with access to health data could
  - Guide to preventative health services
  - Produce a quick summary to health care professionals
- This direction was found promising by interviewed experts

ATOSTEK

Why AI is needed with National Health Records?

- Medical Risk Calculators would be a good technical basis for use of health data, however
  - Often the data does not contain all required inputs
  - Information can be stored in various formats that do not allow using it directly
  - The data may have a lot of variation as it is created with many different information systems by a large number of users

ATOSTEK

Ethical aspects

- Health data can be very personal and must be kept confidential
  - How to prevent the data being used against you?
- In practice, what type of software would be acceptable on the level of a society?
  - Could decisions made by AI system lead to unequal access to health care services?

ATOSTEK

Rules and regulations

- Finland has strict regulations to keep medical data safe
  - The Act on the Secondary Use of Health and Social Data (552/2019) allows researchers access to pseudonymized data in Kanta
  - Product development with statistical data, not with data from individuals
- Finnish Social and Health Data Permit Authority – Findata
  - Permit process can be long
  - All data request must be well justified
  - Data Protection Impact Assessment (DPIA) may be needed
  - Collecting and pseudonymization of data takes time
  - New processes have been created due to the pioneering nature of Jasmine

ATOSTEK

Accessing the data in practice

- The data is made available in Kapseli environment
  - A virtual machine run and operated by Findata
  - Only accessible to named researchers
  - No internet access in the environment
  - Moving any data in and out of the Kapseli happens via Findata
    - How to review learned ML models?
- These issues have an impact on the daily research work

ATOSTEK

Outcome so far

- A delay of more than a year affects plans for work and human resources
- Work was started with other more accessible medical data sets
  - Natural Language Processing methods seem promising with textual data
    - E.g. does a person smoke?
- We have done a review of risk calculators versus Kanta data based on specifications and expert assessments
  - Some relevant data may not be available at all from Kanta
    - E.g. medical history of family members
  - By improving the availability of a few risk calculator parameters in Kanta, we can calculate multiple different risk calculators
  - pre-print: https://papers.ssrn.com/sol3/papers.cfm?abstract_id=4181368

ATOSTEK

Open research questions

- How much data people have in Kanta?
  - Having little or no data for a person limits possibilities
- How much variation there is in recording medical conditions to Kanta?
  - Quality the data, structured data vs. non-structured data
- What is the quality of the textual data?
  - Usefulness and accuracy of the written notes?
- What is the process of updating ML models with new data from Kanta?
  - Required to keep up with changes in structure of Kanta data, used natural language, public health, etc.

ATOSTEK

Secondary use of health data in Kanta for research

Klaus Förger, *D. Sc. (Tech.)*
klaus.forger@atostek.com

www.atostek.com